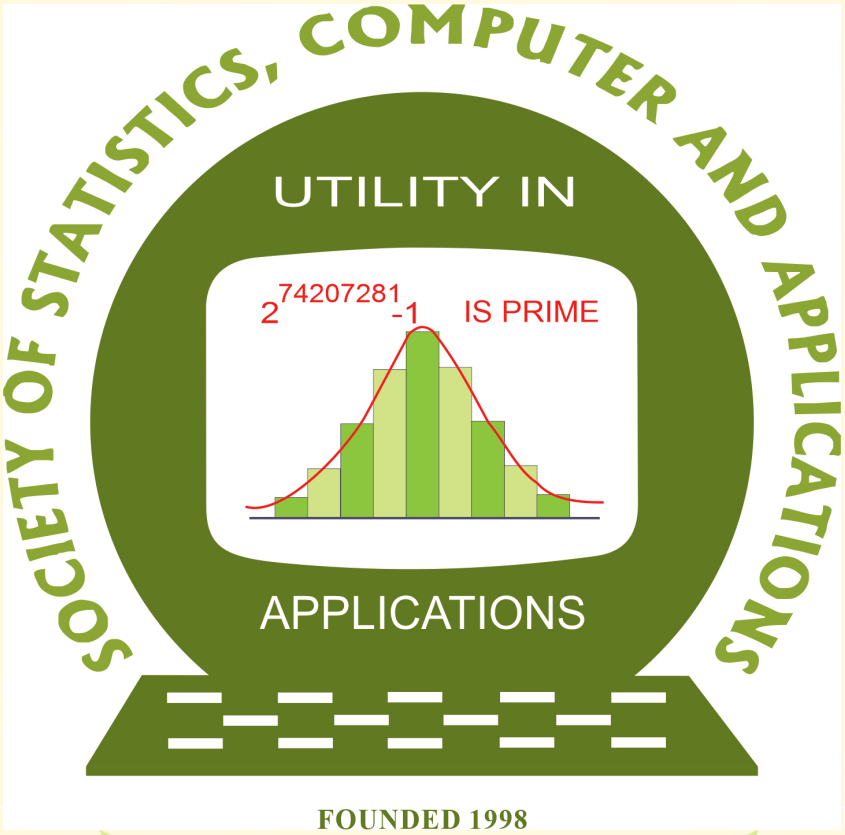
ISBN 978-81-950383-3-6

Special Proceedings (28)
(Based upon the 28th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2026 held at the Department of Statistics and Centre for Advanced Studies, Savitribai Phule Pune University, Ganeshkhind, Pune-411007, India, during January 5-7, 2026)



Society of Statistics, Computer and Applications
https://ssca.org.in
2026

Society of Statistics, Computer and Applications

Council and Office Bearers

President-in-Chief

Late M.N. Das (Founder) Vinod Kumar Gupta

Presidents

Ashish Das Rajender Parsad

Executive Presidents

Jyotirmoy Sarkar Manish Sharma Manisha Pal

Patrons

A.C. Kulshreshtha A.K. Nigam Bikas Kumar Sinha D.K. Ghosh
K.J.S. Satyasai Pankaj Mittal Prithvi Yadav R.B. Barman
R.C. Agrawal Rahul Mukerjee

Vice Presidents

A. Dhandapani P. Venkatesan Praggya Das Ramana V. Davuluri
V.K. Bhatia

Secretary **Foreign Secretary**
Vishal Deo Abhyuday Mandal

Treasurer

Ashish Das

Joint Secretaries

Bishal Gurung Raakhi Singh Shibani Roy Choudhury

Council Members

Aloke Lahiri B. Re. Victor Babu Banti Kumar Dipak Roy Choudhury
Imran Khan Mukesh Kumar P.K. Das Piyush Kant Rai
Rajni Jain Ramasubramanian V. Raosaheb V. Latpate Shalini Chandra
Sukanta Dash V.M. Chacko Vishnu Vardhan R.

Ex-Officio Members (By Designation)

Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Editor-in-Chief, Statistics and Applications
Chair Editors, Statistics and Applications
Executive Editors, Statistics and Applications

Society of Statistics, Computer and Applications
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, India

ISBN 978-81-950383-3-6

Special Proceedings (28)
(Based upon the 28th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2026 held at the Department of Statistics and Centre for Advanced Studies, Savitribai Phule Pune University, Ganeshkhind, Pune-411007, India, during January 5-7, 2026)

Editors
Rajender Parsad
Ashish Das
Durba Bhattacharya
Sourish Das
Baidya Nath Mandal
Vishal Deo

Copyright ©Institutional Publisher: Society of Statistics, Computer and Applications, New Delhi - 110019
Date of Publication: October 25, 2026

Published By:
Institutional Publisher: Society of Statistics, Computer and Applications
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, India
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2026

Printed by: Galaxy Studio & Graphics
A-181, Weavers Colony, Ashok Vihar, Phase-IV, New Delhi-110052
Mob: +91 98183 52203, +91 95829 41203
Email: galaxystudio08@gmail.com

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026

CONTENTS

	Preface	i-ii
1.	Recasting Economic Measurement: The Promise and Trade-offs of Alternative Data *Ajit R. Joshi and Shweta Kumari*	1–12
2.	Stochastic Models for Epidemic Processes in Open Populations with External Infections *Harsha Madushani U. K. A., Christine Soteros and Raj Srinivasan*	13–36
3.	Professor VK Gupta's Influence on Design of Experiments and Sample Surveys *Rajender Parsad*	37–62
4.	Enabling Artificial Intelligence in Health: Perspectives on Health Data Challenges, Mitigation Strategies, and the Current Ecosystem in India *Vishal Deo and Kratika Pandiya*	63–74
5.	Waiting Time Until Success *Jyotirmoy Sarkar*	75–95
6.	CPI Housing Index Methodology *Ashish Das and Praggya Das*	97–113
7.	Performance Assessment of Statistical, Machine Learning and Hybrid Forecasting Approaches for Pearl Millet Production in the Jammu Division *Divyam Sharma, Manish Sharma and Nishant Jasrotia*	115–128
8.	Estimating the Hazard Functions of Two Alternating Recurrent Events *Sugata Sen Roy and Moumita Chatterjee*	129–141
9.	Valuing Free Digital Services in India: Experimental Consumer-Welfare Estimates and the 2025 SNA *Gulab Singh and A. C. Kulshreshtha*	143–168
10.	Virtual Digital Assets: The Need for Regulatory Framework in India *Praggya Das*	169–190
11.	Financial Statistics in India: Current Contours and Future Challenges *Partha Ray*	191–200
12.	Shrinkage Priors for Bayesian Inference – A Contextual Survey *Anandamayee Majumdar*	201–218

13. Foundations First: How Will Mathematics and Statistics Shape India's AI Ambitions? 219–226
Sudeshna Guha Neogi, Abhijit Kulkarni, Ramasubramanian Sundararajan, Balaji Raman and Sourish Das

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026

PREFACE

We are pleased to present the Special Conference Proceedings 2026 from the twenty eighth Annual Conference of the Society of Statistics, Computer, and Applications (SSCA), an International Conference on Innovative Trends in Statistical Science, Data Analytics and Interdisciplinary Research, organized by the Department of Statistics and Centre for Advanced Studies, Savitribai Phule Pune University, Ganeshkhind, Pune, India from January 5-7, 2026.

Founded in 1998, with its inaugural gathering at Haryana Agricultural University, Hisar, SSCA has consistently organized annual conferences at various educational institutions across the country. The society's core mission is to foster research at the intersection of Statistics and Information Technology, supporting both theoretical and applied statisticians dedicated to technological advancements for societal progress. SSCA also promotes open access to knowledge through its journal, Statistics and Applications, which facilitates free downloads, saving, and printing of full-length papers. In addition to regular issues, the journal periodically releases special volumes addressing significant global and national themes.

The recent 28th conference aimed to provide a unified platform for discussions on regional and global statistical issues. Distinguished experts in theoretical and applied statistics from India and abroad, particularly from the USA, actively participated in knowledge dissemination. Speakers represented prestigious Indian institutions, including the Indian Statistical Institute, Indian Institute of Technology (IITs), Indian Council of Agricultural Research (ICAR), Reserve Bank of India (RBI), Universities, and Government offices. The conference featured several notable events, including a Panel Discussion and various technical sessions. These sessions included the M. N. Das Memorial Lecture and a dedicated session on Financial Statistics, where renowned statisticians and leading practitioners shared insights on finance-related topics. Additionally, four endowment lectures were presented : the B. K. Kale Memorial Endowment Lecture, the J. K. Ghosh Memorial Endowment Lecture, the Bikas Kumar Sinha Endowment Lecture and the V. K. Gupta Endowment Award Lecture for achievements in statistical thinking and practice. These lectures were delivered by speakers closely associated with, or students of, the respective honorees. Furthermore, the Aditya Shastri Memorial Lecture in memory of the late Aditya Shastri, Vice Chancellor of Banasthali Vidyapith, was also presented.

The Executive Council of SSCA resolved to compile the "Special Conference Proceedings 2026," highlighting selected presentations, including those from the specialized Financial

Statistics session. Although constraints limited the inclusion of all invited papers, esteemed speakers were invited to submit their research contributions. Following a rigorous review process, a distinguished selection of 13 papers was accepted for publication in the special proceedings. We extend our sincere gratitude to all authors for the prompt submission of high-quality research papers. We owe a special debt of thanks to all the reviewers whose dedicated efforts ensured a swift and thorough review process, completing it within a short timeframe. Special acknowledgment is also due to all members and office bearers of the SSCA Executive Council for their steadfast support. Furthermore, our deepest appreciation goes to all the family members and students who sponsored the Endowment Lectures. The special proceedings from these sessions have been assigned the ISBN #: 978-81-950383-3-6.

We are confident that the contents of these special proceedings will be immensely beneficial to our readership, fostering further insights and advancements in the field of Statistics and its applications in AI and beyond. We welcome any suggestions for improving future conferences and special proceedings as we continually strive to better serve the statistical community.

Rajender Parsad
Ashish Das
Durba Bhattacharya
Sourish Das
Baidya Nath Mandal
Vishal Deo

October 06, 2026

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 1-12

Recasting Economic Measurement: The Promise and Trade-offs of Alternative Data

Ajit R. Joshi* **and Shweta Kumari**

Department of Statistics and Information Management
Reserve Bank of India

Received: 09 March 2026; Revised: 17 June 2026; Accepted: 18 June 2026

Abstract

Central banks today operate in an environment where information needs are rapidly evolving, requiring indicators that are agile, diverse, timely, and capable of providing a comprehensive view of emerging economic developments. The growing availability of non-traditional data sources, coupled with advances in data science, offers significant opportunities to enhance the economic analysis. Developing new indicators from alternative data sources and applying advanced analytics requires careful evaluation to ensure their reliability and efficacy with established macroeconomic measures. Such evaluation is inherently complex, as existing benchmarks may be insufficient or missing, contain coverage gaps, vary in timeliness, and sometimes lack conceptual alignment, thereby undermining the ability to compare indicators accurately. In this context, this paper provides an overview of emerging indicators, explores potential measurement differences, and examines the challenges of integrating non-traditional data with traditional indicators.

Key words: Alternative data; Machine learning; CPI inflation; Sentiment analysis; Remote sensing; Uncertainty; Economic policy; Geospatial modelling; NDVI; EPU; NLP; Online prices; Agriculture.

AMS Subject Classifications: 62K05, 05B05

*This is an extended version of the keynote address delivered at the 28th Annual Conference of the Society of Statistics, Computers and Applications, co-hosted by Savitribai Phule Pune University, on January 5, 2026. Ajit R. Joshi is the Executive Director and Shweta Kumari is Assistant Adviser in the Department of Statistics and Information Management of Reserve Bank of India. The views expressed in this paper are those of the authors and not the institution to which they belong.

1. Background

Measurement of variables is a fundamental aspect of statistical science and the robustness of the methodology employed in measurement determines the utility of any analysis

Corresponding Author: Ajit R. Joshi
Email: arjoshi@rbi.org.in

carried out using the most advanced techniques. As technology is now part of any activity, the transactions and interactions generate a large quantity of data, without any specific data collection efforts. These have come to be categorised as alternative data, as they are often not part of official statistical methodology, at least thus far. But these alternate data sources are becoming mainstream, and all the policy makers and official agencies are embracing them. In this context, I thought it appropriate to place before you some of the observations while using these data for policy analysis.

For a central bank, which plays the roles of data producer and data user, interest in the so-called alternate data sources emanate from both sides. Globally central banks are taking keen interest in these data and the data science techniques to leverage these rich repositories in their own analysis and to enrich the statistical framework.

Developing new indicators from alternative data sources and applying advanced analytics requires careful evaluation to ensure their reliability and the ability to better track the concepts being measured. Such evaluation is inherently complex, as existing benchmarks may be insufficient, missing, contain coverage gaps, or vary in timeliness. The conceptual alignment of the new measures with the standard methodologies also needs to be understood, so as to contextualise the outcome of the new measures and their comparability with the traditional indicators. Therefore, we propose to provide an overview of a few emerging indicators, explore potential measurement differences, and examine the challenges of integrating non-traditional data with traditional indicators. Considering their distinct dynamics, we seek to deliberate on how they could be treated as reliable source of information in a long-term setting, extending beyond one-time research application.

2. Online food prices - why online prices are important

I will begin with an indicator which is probably the most well-known and central to my work in the Reserve Bank. Thanks to the transparent approach of the Ministry of Statistics and Programme Implementation, we all know that the major price indexes are under revision, and we should see the new ones by next month. In the context of the revision exercise, online prices have emerged as a potential data source because they provide high-frequency, real-time information on the movement of prices. Traditional price collection surveys are typically monthly and rely on field visits, whereas online data can be harvested daily from multiple retailers. This enables analysts and policymakers to detect inflationary pressures before availability of the official statistics. In periods of sudden shocks, such as pandemics, supply-chain disruptions, or major policy changes (for example, the recent GST rationalisation), online prices can reveal changes immediately, well before these appear in official inflation measures. Their speed, wide product coverage, and relatively low data collection cost make them an important tool in the price-monitoring ecosystem. For economies with expanding e-commerce, online prices increasingly reflect the prices paid by a large share of consumers. Thus, they are innovative, complementary, and valuable for understanding market dynamics.

2.1. Why online prices may not match official inflation numbers

Even though online prices are useful in tracking emerging trends, the indices compiled using them may often diverge from official inflation rates. This can occur due to several structural differences, as the online and offline retail markets are not identical. There can

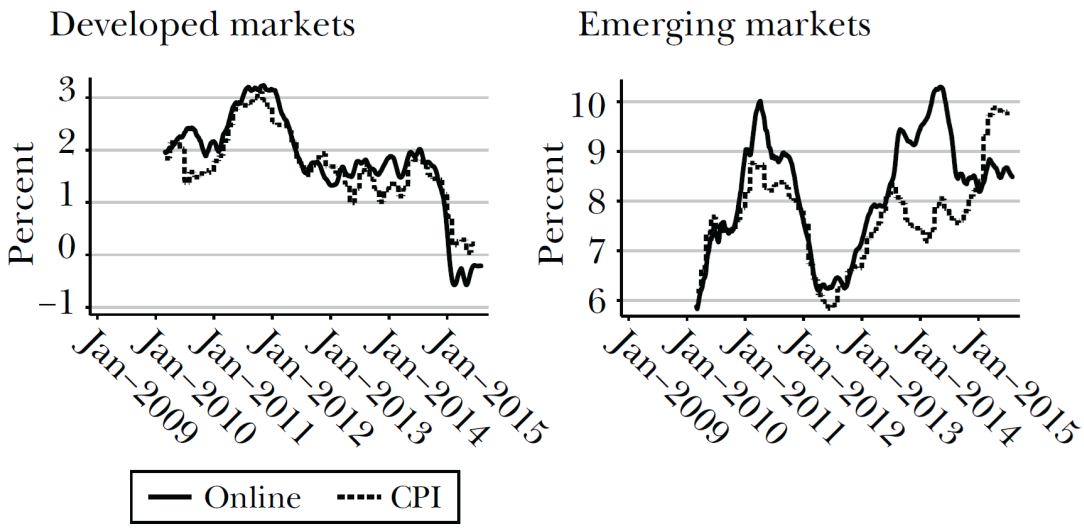


Figure 1: Online v/s CPI - annual inflation rates

Source: Cavallo and Rigobon (2016)

be differences in the coverage of items, the types of commodities sold, and regional price variations. In many places, especially smaller cities and semi-urban areas, a large share of essential daily consumption items is still purchased offline, meaning online prices capture only a partial view. Both offline and online measures seek to capture inflation as a concept, and any divergence between them stems from measurement differences, and should not be treated as an error.

In their Billion Prices Project, Cavallo and Rigobon (2016) find that while the dynamics of online prices is distinct in many aspects from offline sources, the carefully curated online index is able to anticipate movements in the corresponding official consumer price index in many countries. Further, the coherence between the two price indices appears to be more in developed countries, as compared to emerging market economies (Figure 1). While these insights reflect trends observed between 2009 and 2015, current patterns may have evolved significantly after a decade.

2.2. Why we should not target to match with offline prices - but still value them

Given these structural differences, online inflation indices should not be expected to precisely match official CPI estimates. They are designed to track fast-moving price trends, rather than measurement of cost-of-living changes. Their underlying market dynamics, such as the degree of price stickiness, can differ substantially from offline retail markets, leading to different patterns of price adjustments. For example, online competition may lead to frequent changes in prices, as Cavallo (2018) found in the case of US for large multichannel retailers. Price setting can further be influenced by automated or algorithmic pricing mechanisms used by online retailers, as noted by Brown and MacKay (2023), unlike offline retail prices, which are typically revised at discrete and infrequent intervals. In contrast, the official CPI adheres to international statistical standards, reflect representative household expenditure weights,

and follow consistent long-term methodology - requirements that online datasets may not meet. This distinction, however, reinforces rather than diminishes the value of online price data. Their strength lies in offering an additional perspective on price dynamics, which is useful for early detection of market conditions, stress, or structural shifts.

As examined by Mahajan and Tomar (2021) in the Indian context, supply-chain disruptions triggered by pandemic lockdowns resulted in reduced product availability in both online grocery portal and offline markets (mandis). However, the magnitude of decline was found to be lower in online market for select food categories.

Such findings provide deeper insights into how crisis episodes alter supply-side mechanism, price-setting behaviour, and market dynamics, extending beyond the usual focus of co-movement between online and offline prices. Importantly, such nuances are difficult to detect using conventional offline price data alone, and underscore the value of online prices data further, where prices and availability are recorded at high frequency and responses are faster.

Online prices, therefore, function as a complementary information stream that enhances situational awareness and supports research. Their purpose is not to be a substitute or completely mirror CPI, but to provide insights that enrich monitoring and analysis.

3. Remote sensing - policy perspective

Climatic factors have emerged as an important risk area, drawing heightened attention across global policy forums in recent years. While climate risk is a very broad domain, the inter-linkages with macro-economy and the financial sector are important for the Reserve Bank. Here we focus only on the impact of the weather events on crop conditions and resultant effect on the production. We are attempting evaluation of the supply-side situation using the real time data on weather conditions, and satellite imagery which provide early signals on crop growth conditions and identify the significant stress factors.

3.1. Agricultural data challenge

Data on crop production comes with considerable delay and there are gaps in crop area and production estimates, particularly at disaggregated regional levels. This task is inherently complex considering the vast landscape of our country, diversity of crops, and varying production cycles. The traditional system relies on field surveys and reporting cycles, which limits the availability of timely updates during the crop season. Crop production estimates are also revised subsequently, and the revision is often significant, constraining real-time assessment of evolving supply conditions. The intricacies of agricultural ecosystems get further amplified by frequently changing climatic conditions. As a policy institution, we need information as of yesterday and that is where we started looking up the sky to see whether the satellite data can help us.

3.2. What makes remote sensing a powerful data source?

Remote sensing offers a distinct advantage in agricultural monitoring by enabling regular, large-area observations of crop lands in a consistent and systematic manner (Donaldson

and Storeygard, 2016). Satellite-based measurements facilitate objective assessment of crop health across regions and seasons, reducing dependence on field observations. Importantly, remote sensing captures in-season crop dynamics - such as sowing progress, phenological stages, vegetative growth, and emerging stress - thereby generating timely signals on evolving crop conditions, often well before their full impact is reflected in traditional data (Gumma *et al.*, 2022; Johnson, 2014). This capability is particularly valuable for India, where spatial and temporal variability in climatic condition is substantial.

Geospatial capabilities and change-detection frameworks based on satellite imagery enable early assessment of crop stress and damage arising from droughts, floods, heatwaves, and other extreme weather events. Such "pixels to plots" insights strengthen early warning systems, bridging critical information gaps during the crop season. From sowing to harvesting, remote sensing allows for periodic updates, and production nowcasts well ahead of traditional agricultural statistics.

A recent study by Kumari (2024) provides a scalable framework for integrating remote sensing data for regular crop monitoring and production assessment. Figure 2 illustrates how such data capture spatio-temporal variation in vegetation conditions for Tur and Gram - two major pulses in India - across comparable fortnights in key growing regions. The 2016-17 crop season, which recorded a strong Tur harvest, exhibits higher vegetation levels as early as August 2016, relative to August 2015. Similarly, for Gram, which is a Rabi crop, improved conditions are reflected in November 2021 compared to November 2015.

Importantly, the vegetation index adds substantial value by enhancing both the accuracy and timeliness of production assessment, enabling real-time nowcasts, with estimates getting progressively refined and finalised near harvest phase. The study shows that model-based predictions closely track both upturns and downturns in actual production data.

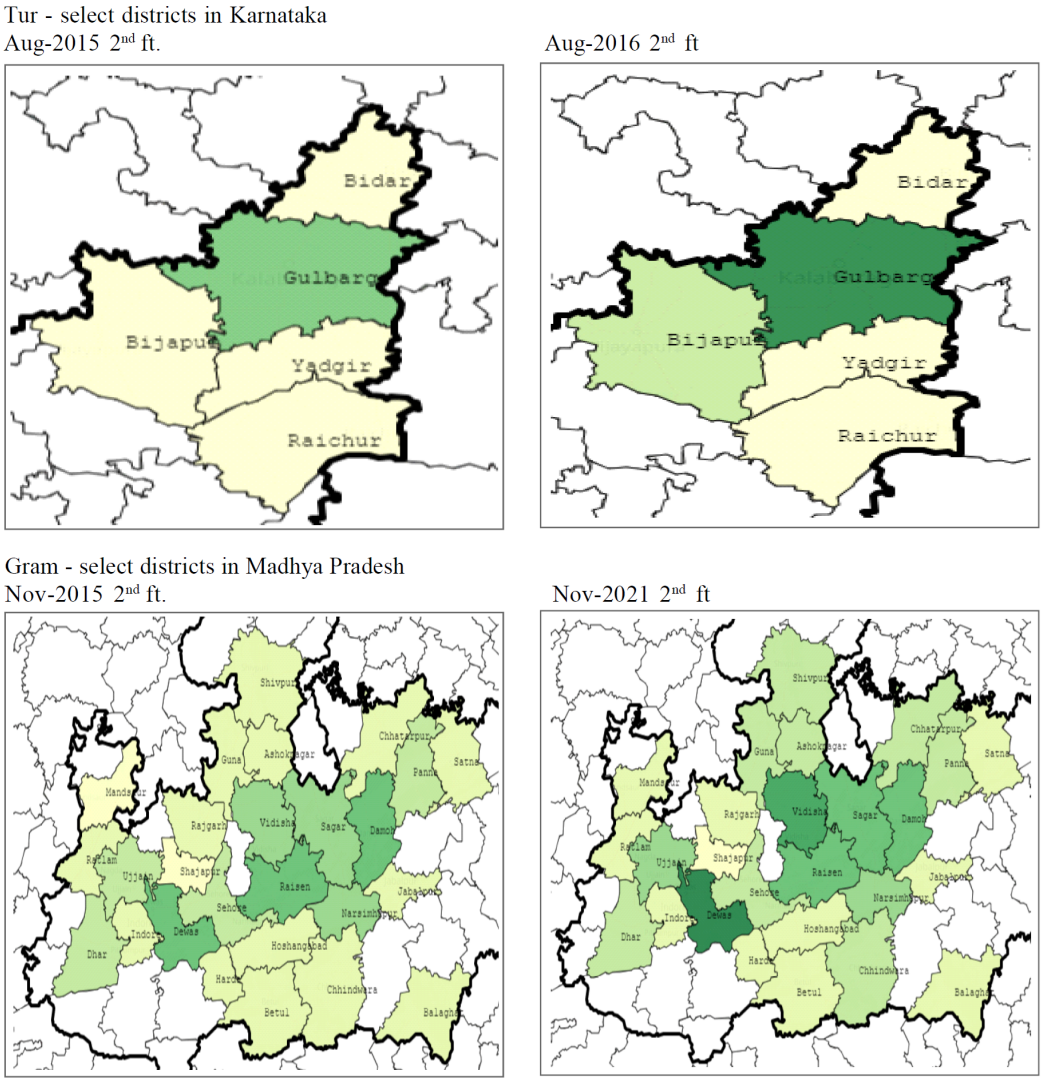
Further, vegetation conditions are a rich source of advance information for assessing agricultural commodity arrivals in mandis, which, in turn, can influence the future trends in wholesale and retail prices. In the Indian context, Kumari and Kuruganti (2022) combine remote sensing data with ground-based indicators to predict mandi arrivals, demonstrating the usefulness of such information.

This study also captures spatial heterogeneity in vegetation-arrival relationship across mandis, using geographically weighted regression (GWR) model (Figure 3). This aspect is difficult to identify using traditional data sources, but it is possible by leveraging the spatial granularity of remote sensing data through explicit incorporation of geolocations.

It is learnt that remote sensing-based inputs are increasingly used internally to inform official crop estimates (Ministry of Agriculture and Farmers Welfare (MoA&FW), 2025; Navalgund and Ray, 2019); however, their specific role and quantitative contribution are not published separately.

3.3. Why the difference

While remote sensing offers high-frequency, real-time spatial observations, and its ability to track actual crop production is well-established in various research studies, its estimates can sometimes differ from traditional crop statistics because of structural and



Note: Yellow to Green gradient scale represents lower to higher vegetation value.

Figure 2: Geospatial variation of vegetation

Source: Kumari (2024)

measurement differences. The indicators - such as vegetation indices, soil moisture, photosynthetically active radiation (PAR) - are biophysical proxies, rather than direct measures of crop yield or production. Differences also arise from scale and granularity. Satellite data record information at the pixel level and can cover heterogeneous plots, capturing fine spatial variation that is often missed in aggregated data. Because official statistics are aggregated at broad administrative levels, downscaling granular satellite data to these levels can reduce its informational richness and contribute to divergence.

Timing is another factor, satellites provide continuous in-season insights, whereas field surveys are periodic and may lag by weeks or months. At high-frequency or local levels, corresponding official data is not available, as national statistics typically provide a single estimate for the whole crop year (which gets updated subsequently). By the time state or district level estimates are published (often 1-2 years lag), timely comparison becomes

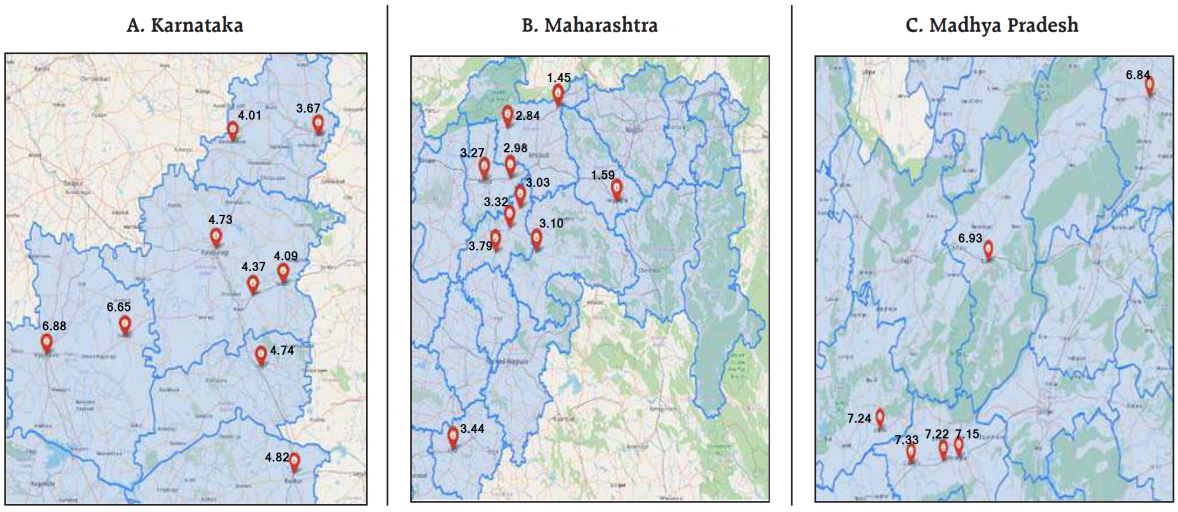


Figure 3: Vegetation coefficient at mandi level - GWR model

Source: Kumari and Kuruganti (2022)

impossible. Methodological variations - including crop classification and phenological models, area selection frameworks, weighting patterns, and data calibration techniques - further influence the assessment.

3.4. Way forward

The differences in the two sets of estimates are to be expected. Understanding them helps interpret remote sensing outputs appropriately and leverage the rich dataset. While ground surveys remain useful for validating yields, satellite-based monitoring adds a dynamic, high-resolution complementary perspective. In this way, both these techniques can help the other data collection efforts more comprehensive and provide opportunity for mutual learning. As analytical and geospatial techniques advance and data integration improves, the value of remote sensing for agriculture is expected to grow further. There should be no expectation for remote sensing data to perfectly match traditional statistics. Instead, by combining 'boots on the ground' with 'eyes in the sky,' a more nuanced and holistic view of agricultural conditions can be obtained. This integrated perspective would go a long way in providing actionable intelligence to manage climate risks and their impacts on macro-economy.

4. Soft data – why we need more than just hard data

Monitoring the economy requires going beyond conventional statistical indicators *viz.* inflation, growth, production, trade or employment. While these "hard numbers" remain indispensable, they capture only the observable outcomes of economic activity, and do not fully reflect different types and levels of disaggregation, evolving expectations, and behavioural responses that significantly influence economic dynamics. Soft data, comprising media reports, expert analysis, sector-specific studies, and corporate communication, helps capture these dimensions.

4.1. Media as a bridge

Media acts as a crucial communication bridge between households, markets and policymakers by reporting and reflecting the perceptions and expectations of economic agents. Even in periods when official statistics or surveys face constraints, as seen during the COVID-19 pandemic, such soft information remains accessible, offering timely insights into economic conditions.

4.2. Early signals

A major strength of news-based information is its timeliness. Many macroeconomic indicators are released with lag, revised later, or available only at low frequency. Media reports, on the other hand, often pick up emerging pressures or on-ground reality much earlier. Coverage of rising input costs, supply chain disruptions, shifts in consumer demand, rural distress, trade bottlenecks, or early signs of recovery provide valuable real-time signals well before these are captured by the official statistics. In this sense, soft data serves as an important source for an early-warning system.

Relationship of news-based sentiment indicators with economic activity, either alone or in combination with other economic indicators, has been examined on a range of variables in many countries and is increasingly being used in forecasting frameworks. Shapiro *et al.* (2017) construct monthly news sentiment index for US from a wide range of economic and financial news articles for a long period from 1980 to 2015, and find a strong correlation with key economic variables. Further, their predictive power is also evaluated and found to be significant in current and near-future term.

Another study in the context of US by Ardia *et al.* (2019) develops a comprehensive design for optimisation of news-based sentiment. Use of sentiment is found to provide valuable gains for forecasting economic growth, relative to forecasts obtained using only the economic and financial indicators (hard data).

Kalamara *et al.* (2020) extract signal from UK newspapers and find that it significantly improves the forecast of key macroeconomic variables such as GDP, inflation, and unemployment, as compared to benchmark model, emphasising the valuable information content.

Various empirical studies indicate that soft information offers key insights, and are particularly useful during periods of stress or economic uncertainty (Aguilar *et al.*, 2021; Buckman *et al.*, 2020). In the Indian context, Kumari and Giddi (2020) derive sentiment from news using machine learning model (Support Vector Machines classifier) and find strong linkage and causality between the sentiment index and retail inflation.

Further, Giddi *et al.* (2022) construct news based sentiment indices for a variety of macroeconomic variables using a fine-grained aspect based sentiment methodology. They find that pessimism was reflected in news narratives much before the impact was recorded in hard data, whose values significantly deviated from historical trends, especially during the first phase of Covid-19. The sentiment indices capture the changing patterns in various sectors of Indian economy and have significant predictive value.

These findings suggest that a high-frequency sentiment index can be a useful com-

plementary indicator to provide early signals on economic conditions. However, considering that textual reports could be noisy, it is important to process the soft information carefully in order to capture the underlying nuances and extract useful signals. Corporate news is another crucial component, particularly for large firms which often act as market makers in pricing, investment, production and employment, and influence wider economic outcomes. Analysing corporate reports, such as earnings calls, management commentary or interviews, and analysts' reports help quick assessment of potential turning points in business cycles.

While analysts often review multiple information sources manually, they often put their analysis out in the public domain or share with policy makers. An automated pipeline can systematically collate and analyse narratives, extract sentiment and key insights, and present the emerging trends in a structured manner. Using NLP and ML techniques, the tasks can be suitably streamlined in a step-wise manner from input data source to the desired output formats, enabling policymakers to quickly decipher the perspectives, enhancing the overall efficiency and consistency, while reducing the time and effort required for manual review of large textual documents.

4.3. Explaining the "Why" behind economic movements

Soft information adds analytical depth by helping explain the "why" behind movements in macroeconomic variables. While hard data may show inflation rising or industrial output slowing, news coverage helps identify the underlying reasons - such as supply constraints, global demand shifts, policy uncertainty, financial conditions, or changes in business and consumer sentiment. These data provide the contexts which are not captured in hard data, and enable analysts to interpret emerging trends more accurately, going beyond quantitative measures in isolation.

4.4. Why uncertainty matters

Uncertainty often shapes the decisions of economic agents (how companies invest, how households spend or save, how markets behave) and influences how policymakers respond, creating a feedback loop that further affects economic outcomes. While the idea of uncertainty is well recognised, its measurement is far from straightforward. Uncertainty is an intangible, multi-dimensional concept, and unlike inflation or growth, it does not have a directly observable counterpart.

4.5. Relevance of text-based indices

Beyond financial market volatility indicators and business-survey-based measures, recent years have seen the emergence of soft-data based measures of uncertainty. In this context, text-based uncertainty indices such as the Economic Policy Uncertainty (EPU) Index, Geopolitical Risk Index (GRI) and Trade Policy Uncertainty (TPU) have gained remarkable traction (Baker *et al.*, 2016; Caldara and Iacoviello, 2022; Caldara *et al.*, 2020). By covering a broad spectrum of issues from fiscal and monetary policy developments to geopolitical tensions, trade disputes and tariff actions, they offer a comprehensive view of evolving risks, many of which may not immediately appear in traditional macro variables. These indices provide timely, nuanced, and quantitative signals of uncertainty by tracking emerging media narratives on economic policies, trade, and geopolitical risks. These text-

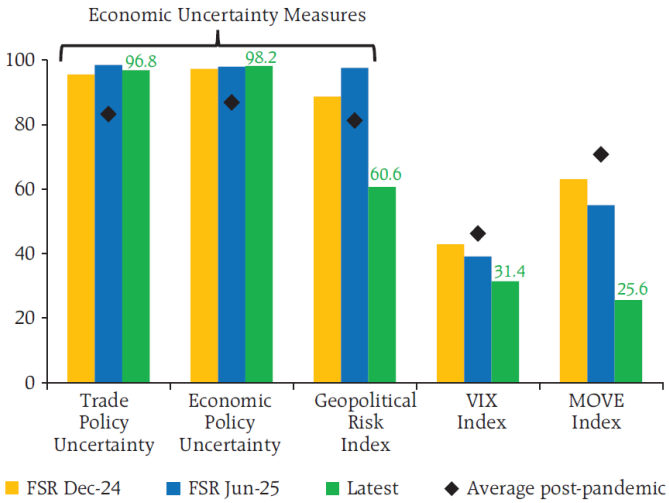


Figure 4: Disconnect between uncertainty and financial market volatility (percentile)

Source: For details, please refer to Financial Stability Report (FSR) Dec 2025, RBI

based uncertainty indicators are often analysed jointly with financial market indicators (*e.g.* VIX, MOVE index) for a comprehensive assessment of evolving economic conditions (Ahir *et al.*, 2025; Baker *et al.*, 2020).

This is particularly relevant for a dynamic and fast-growing economy like India, where policy decisions and evolving developments (including potential external shocks) can quickly influence expectations and market confidence. Their high-frequency nature makes them particularly valuable during fast-moving episodes such as tariff tantrums, geopolitical escalations, or major policy shifts. Their standardised methodology also enables meaningful cross-country comparison, supporting analysis of global uncertainty dynamics and potential spillovers, which are significant for EMEs. It is, therefore, important to monitor a range of indicators on a regular basis to detect any signs of emerging stress and vulnerabilities (Economic Survey 2025-26; FSR 2025; WEO 2026). These indicators should, however, be used with caution, as different indicators may not always converge and convey the same degree of uncertainty, as seen during recent times (Figure 4).

4.6. Benchmarking

Evaluating the effectiveness of these uncertainty indicators is inherently challenging because there is no single "true" economic indicator against which they can be directly benchmarked. There is also a natural difference in the way these are constructed and interpreted. While heightened uncertainty generally tends to dampen investment and productivity, such relationships typically manifest over the business and financial cycle, rather than in short-term, although the magnitude and persistence of such effects vary across countries (Castelnuovo, 2023).

In addition, causality is difficult to isolate, as uncertainty itself may be both a driver and a reflection of economic stress. News-based indices may also be influenced by reporting

intensity, narrative framing or country-specific contexts. Therefore, while they may not always have a direct, contemporaneous, or mechanical link to real economy, these indicators remain valuable as narrative and complementary tools that enrich economic understanding and policy assessment.

5. Concluding observations

The increasing complexity and digitalisation of economic activity have opened a treasure trove of information for the public policy institutions. This has not only enabled the development of completely new indicators but also created opportunities to complement and enrich conventional statistical measures with more timely, granular, and near real-time insights, which is becoming increasingly necessary. As alternative indicators become more widely used, there is a natural tendency to assess them by how closely they align with traditional measures. However, this framing can be limiting, as it places undue emphasis on correspondence rather than informational value.

The more relevant question is how these indicators help improve understanding of evolving economic dynamics and whether they add insights that are not readily available from existing data sources. The utility of such indicators is therefore best evaluated by examining how they contribute to decision-making within structural and forecasting frameworks. Ignoring the relevant indicators could expose policy analysis to model and data-quality risks. As the saying goes, "the proof of the pudding is in eating"; similarly, the value of new-age indicators will ultimately be established through their performance in practical analytical applications.

Broadening the statistical information base and remaining open to new indicators, while exercising appropriate due diligence, can significantly strengthen the analytical foundations, improve our understanding of economic reality, and bring policy thinking closer to it.

References

Aguilar, P., Ghirelli, C., Pacce, M. and Urtasun, A. (2021). Can news help measure economic sentiment? An application in COVID-19 times. *Economic Letters*, **199**, 109730.

Ahir, H., Bloom, N., and Furceri, D. (2025). Uncertainty about uncertainty. *Finance and Development Magazine*, International Monetary Fund.

Ardia, D., Bluteau, K. and Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, **35**, 1370–1386.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, **131**, 1593–1636.

Baker, S. R., Bloom, N., Davis, S. J., and Terry, S. J. (2020). COVID-induced economic uncertainty (No. w26983). National Bureau of Economic Research.

Brown, Z. Y., and MacKay, A. (2023). Competition in pricing algorithms. *American Economic Journal: Microeconomics*, **15**, 109–156.

Buckman, S. R., Shapiro, A. H., Sudhof, M. and Wilson, D. J. (2020). News sentiment in the time of COVID-19. *FRBSF Economic Letter*, **08**, 5–10.

Caldara, D., and Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, **112**, 1194–1225.

Caldara, D., Iacoviello, M., Molligo, P., Prestipino, A., and Raffo, A. (2020). The economic effects of trade policy uncertainty. *Journal of Monetary Economics*, **109**, 38–59.

Castelnuovo, E. (2023). Uncertainty before and during COVID-19: A survey. *Journal of Economic Surveys*, **37**, 821–864.

Cavallo, A. (2018). More Amazon effects: Online competition and pricing behaviors. National Bureau of Economic Research, Working Paper No. 25138.

Cavallo, A. and Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, **30**,151–178.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, **30**, 171–198.

Economic Survey 2025-26. Government of India, Ministry of Finance.

FSR (Financial Stability Report) (June and December 2025). Reserve Bank of India.

Gumma, M. K., Kadiyala, M. D. M., Panjala, P., Ray, S. S., Akuraju, V. R., Dubey, S., Smith, A. P., Das, R., and Whitbread, A. M. (2022). Assimilation of remote sensing data into crop growth model for yield estimation: A case study from India. *Journal of the Indian Society of Remote Sensing*, **50**.

Giddi, G., Kumari, S. and Maiti, S. S. (2022). When a news story is more than just text: evidence from Indian economy. RBI Bulletin.

Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, **141**.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G. and Kapadia, S. (2020). Making text count: Economic forecasting using newspaper text. Bank of England Staff Working Paper No. 865.

Kumari, S. (2024). From pixels to plots: Synthesizing remote sensing into crop production estimation. *Journal of Income and Wealth*, **46**, 113–123.

Kumari, S. and Kuruganti, S. (2022). Remote sensing applications for policy: An assessment of agricultural commodity arrivals. RBI Bulletin, July.

Kumari, S. and Giddi, G. (2020). Inflation decoded through power of words, RBI Bulletin.

Lobell, D. B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., and Murray, S. (2019). Eyes in the sky, boots on the ground: Assessing satellite- and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, **102**, 202–219.

Mahajan, K., and Tomar, S. (2021). COVID-19 and supply chain disruption: Evidence from food markets in India. *American Journal of Agricultural Economics*, **103**, 35–52.

Ministry of Agriculture and Farmers Welfare (2025). Assessment of crop losses through satellite. Press release dated February 11, 2025 by PIB Delhi.

Navalgund, R. R., and Ray, S. S. (2019). Application of space technology in agriculture: An overview. Smart Agri Post.

Shapiro, A. H., Sudhof, M. and Wilson, D. (2017). Measuring news sentiment. Federal Reserve Bank of San Francisco Working Paper 2017-01.

World Economic Outlook (WEO) update (2026). International Monetary Fund. January.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 13-36

Stochastic Models for Epidemic Processes in Open Populations with External Infections

Harsha Madushani U. K. A., Christine Soteros and Raj Srinivasan

Department of Mathematics and Statistics
University of Saskatchewan, Canada

Received: 29 May 2026; Revised: 25 June 2026; Accepted: 30 June 2026

Abstract

This paper studies stochastic epidemic models for infectious disease transmission in an open population with external infections. We highlight the limitations of classical deterministic epidemic models, particularly the Susceptible-Infected-Recovered (SIR) framework, which assumes a closed population and does not adequately capture the uncertainty and randomness inherent in epidemic spread. Stochastic birth-death models incorporate random fluctuations and provide a more realistic probabilistic description of epidemic dynamics. However, standard birth-death formulations generally do not account for infections entering the population from external sources. To address this limitation, we consider a birth-death process with immigration, where immigration represents infections introduced from outside the population. The paper reviews the unpublished work of Kobayashi, who first proposed this birth-death-immigration framework in the context of COVID-19 modelling. One of the remarkable features of this model is that it admits an exact analytical solution in terms of a generalized negative binomial distribution.

The paper presents the probability distribution, expected values, and variances associated with the model, and discusses two immigration mechanisms: all-state immigration and zero-state immigration. In addition, several extensions involving time-varying infection and immigration rates are investigated to demonstrate how nonhomogeneous assumptions influence epidemic behaviour and intervention outcomes. Since this work is intended as a review and exposition of Kobayashi's unpublished results, the focus is on summarizing the theoretical results, clarifying the probabilistic structure of the model, and discussing possible extensions and applications in epidemic modelling.

Key words: Stochastic models; Infectious disease dynamics; Open populations; External infections; Birth-death-immigration process; Time-nonhomogeneous models.

Corresponding Author: Raj Srinivasan
Email: raj@math.usask.ca

1. Introduction

In recent years, the limitations of purely deterministic epidemic models have become increasingly apparent, particularly in the context of emerging infectious diseases and globally connected populations. Deterministic compartmental models such as the Susceptible-Infected-Recovered (SIR) model (see a review in Allen *et al.* (2008)), although mathematically tractable and useful for large-scale approximations, often fail to capture the stochastic variability that is naturally present in epidemic transmission. Such variability becomes especially important when infection counts are small, when outbreaks begin through isolated introductions, or when external infections continuously reintroduce disease into a population.

Classical stochastic birth-death models improve upon deterministic formulations by incorporating randomness directly into the infection and recovery mechanisms (Bailey, 1991). These models provide insight into extinction probabilities, variability in outbreak sizes, and the occurrence of rare but significant epidemic events. However, many standard birth-death epidemic models still assume effectively closed populations and therefore do not explicitly model infections entering from outside sources.

The present paper focuses on a birth-death process with immigration, where immigration represents external infections entering the population. The model considered here is based on the unpublished work of Kobayashi (2021b,c), who introduced this framework for modelling COVID-19 transmission dynamics. A particularly striking aspect of the model is that it possesses an exact analytical solution characterized by a generalized negative binomial distribution. This feature makes the model mathematically elegant while also providing a useful probabilistic description of epidemic spread in open populations.

The primary goal of this paper is not to introduce a completely new model, but rather to provide a careful review, exposition, and summary of Kobayashi's unpublished results. In doing so, we preserve the original mathematical structure while presenting the results and interpretations in a unified and accessible manner. The paper also discusses possible extensions, including time-nonhomogeneous formulations in which infection and immigration rates vary over time. Additional details, including derivations and numerical results, can be found in Madushani U.K.A. (2025).

Infectious diseases have posed a persistent threat to human health throughout history, leading to significant morbidity, mortality, and societal disruption, with some escalating into devastating global pandemics. Understanding and predicting the spread of these diseases is paramount for implementing effective public health interventions and mitigating their impact. Mathematical models have become indispensable tools in this endeavour, offering frameworks for analyzing disease transmission dynamics and evaluating potential control strategies. The mathematical study of infectious diseases is considered to have originated with John Graunt in 1662 (see a review of the history in Brauer (2017)). Subsequently, as reviewed by Kretzschmar and Wallinga (2010), a foundational model in epidemiology was formulated in 1927 by Kermack and McKendrick: the deterministic Susceptible- Infected-Recovered (SIR) model. This model, based on dividing the population into compartments representing individuals' disease status, marked the beginning of compartmental modelling and provided a basic deterministic framework for understanding disease spread within a closed population. The introduction of the "Basic Reproduction Number" (R_0) by Macdonald (1952) further enhanced the understanding of disease transmissibility. Over time,

various extensions and modifications of the SIR model, such as SIS (Susceptible - Infected - Susceptible), SEIS (Susceptible - Infected - Exposed - Susceptible, and SVIR (Susceptible - Vaccinated - Infected - Recovered), have been developed to account for different disease characteristics and epidemiological scenarios. However, deterministic models operate under simplifying assumptions, such as fixed infection and recovery rates, which ignore the inherent randomness in disease transmission.

A significant limitation of deterministic models is their inability to account for random variations and the unpredictability of real-world epidemics, thus failing to accurately forecast dynamic disease patterns (Andersson and Britton, 2012). To address these limitations, stochastic models were introduced, incorporating the element of randomness and variability in disease transmission. Stochastic models are crucial for reflecting the unpredictable nature of human interactions and disease progression, capturing the uncertainty inherent in real-world epidemics and accounting for chance occurrences that can significantly alter the course of disease spread. While the foundational concepts of stochastic processes emerged in 1963, their application in epidemiological literature became more prevalent starting in 1978 (see a review in Brauer (2017)). In recent years, Allen (2008, 2010) has significantly contributed to epidemiological research, particularly in the analysis of stochastic epidemic models. Here, we explore the application of birth-death and birth-death-immigration stochastic models to understand the spread of disease, particularly in open populations where external infections can occur. The concept of the birth-death process, initially used for population growth analysis, has been adapted to model infectious disease dynamics, focusing on the number of infected individuals over time. In 1939, Feller highlighted the significance of stochastic processes for understanding population growth by analysing a birth-death process where the birth rate λ and death rate μ are constants. Kendall (1948b) extended Feller's work by considering the birth and death rates as functions of time t denoted as $\lambda(t)$ and $\mu(t)$. Since then, various birth-death models have been developed under different assumptions about the temporal dynamics of model parameters. Bailey (1991) and Andersson and Britton (2012) showed that a birth-death process can also be used to study infectious disease dynamics where the number of individuals infected is the main interest. Allen *et al.* (2017) demonstrated ways to use birth-death models in epidemiology to capture the dynamic behaviour of infected individuals in a system. The immigration concept was introduced to the birth-death process by Kendall (1948a) and birth-death immigration processes were reviewed in the book "The Elements of Stochastic Process with Applications to the Natural Sciences" by Bailey (1991). For modelling infectious diseases, "immigration" can be used to represent the introduction of new infections from external sources, herein called external infections. The concept of external infection was not utilized in COVID-19 epidemiology until Kobayashi (2020) discussed the significance of the birth-death-immigration stochastic model for studying open populations. He introduced stochastic models based on the birth-death process with external infection for homogeneous and non-homogeneous cases, incorporating random variations in infection and recovery/death rates in his unpublished series of papers. He noted that these stochastic models could offer more accurate predictions of epidemic patterns in real-world situations (see Kobayashi (2020, 2021a,b,c)).

The research presented here aims to contribute to a deeper understanding of infectious disease dynamics in open populations by: i) (Section 2) Reviewing the deterministic SIR model and emphasizing its limitations, particularly the closed population assumption and the lack of inherent variability; ii) (Section 3) Examining the application of a basic

stochastic birth-death model in epidemiology and discussing its limitations, such as the absorbing state of disease extinction; iii) (Section 4) Investigating birth-death-immigration models under two distinct policies for external infections: all-state immigration (continuous) and zero-state immigration (reintroduction after extinction); and iv) (Section 5) Analyzing the impact of varying internal and external infection rates on the expected number of infected individuals in a time-nonhomogeneous birth-death all-state immigration model, thereby exploring the effects of intervention measures. By addressing these objectives, this paper aims to underscore the importance of stochastic modelling, particularly birth-death-immigration processes, for gaining valuable insights into the uncertainties and variations inherent in real-world epidemic scenarios.

2. Review of deterministic SIR model

The deterministic SIR model serves as a foundational framework in epidemiology for understanding the spread of infectious diseases within a population. Despite having limitations, it provides essential insights into the basic dynamics of epidemics. The deterministic SIR model divides a population of size N into three compartments: Susceptible (S), Infected (I), and Recovered (R). It assumes a closed population where the total size remains constant $S(t)+I(t)+R(t) = N$. The model tracks the flow of individuals between these compartments based on two key parameters:

- β (Per-capita infection rate): Represents the rate at which susceptible individuals become infected through contact with infected individuals. The rate of new infections is proportional to the product of the number of susceptible and infected individuals $S(t)I(t)$.
- γ (Per-capita recovery rate): Represents the rate at which infected individuals recover from the disease and become immune. This is the inverse of the average infectious period.

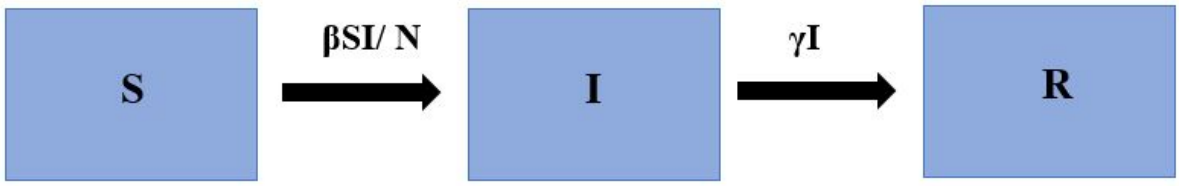


Figure 1: SIR compartmental diagram: Susceptible (S), Infected (I), Recovered (R):
$S(t) + I(t) + R(t) = N$ **for all** t**.**

The dynamics of the deterministic SIR model are described by a system of ordinary differential equations:

$$\frac{dS(t)}{dt} = -\frac{\beta S(t)I(t)}{N} \frac{dI(t)}{dt} = \frac{\beta S(t)I(t)}{N} - \gamma I(t) \frac{dR(t)}{dt} = \gamma I(t).$$

These equations describe the rates of change in the number of individuals in each compartment over time.

Two critical quantities derived from the deterministic SIR model are the Basic Reproduction Number R_0 and the Effective Reproduction Number R_t.

- **Basic Reproduction Number** R_0: Defined as the average number of secondary infections generated by a single infected individual in a completely susceptible population.
$$R_0 = \frac{\beta}{\gamma}.$$
- **Effective Reproduction Number** R_t: Represents the average number of secondary infections produced by an infectious individual at a specific time t.
$$R_t = R_0 \cdot \frac{S(t)}{N}.$$

The **Final Epidemic Size** in the deterministic SIR model refers to the total proportion of the population that becomes infected and recovers by the end of the epidemic. Let s_∞ be the proportion susceptible at the end and let r_∞ be the proportion recovered at the end, then:
$$s_\infty = e^{R_0(s_\infty - 1)}, \quad r_\infty = 1 - s_\infty = 1 - e^{-R_0 r_\infty}.$$

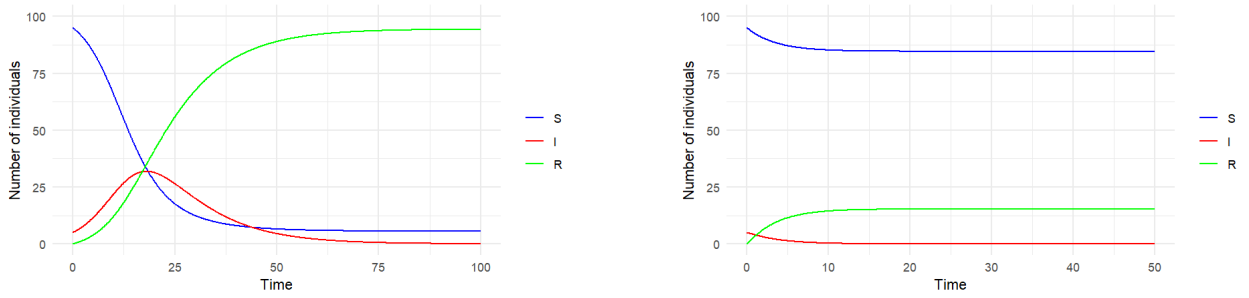
2.1. Example numerical results for the SIR model

In order to compare later to birth-death-models, this section illustrates a numerical solution of the deterministic SIR model. Euler's method was used to explore the dynamic behaviour of each compartment: Susceptible, Infected and Recovered, considering two different scenarios of R_0: $R_0 < 1$ and $R_0 > 1$. This approach allows us to computationally simulate and analyze infectious disease transmission dynamics within a closed population under specific model parameters. We select a population $N = 100$ and the initial number of infectives $I(0) = 5$ for both cases. Thus, $S(0) = 95$ and $R(0) = 0$. Two figures for $R_0 < 1$ and $R_0 > 1$ are generated for two different sets of parameters to illustrate the dynamic behaviour of each compartment.

In Figure 2(a), the infection rate $\beta = 0.3$ (i.e., on average, each infected individual will transmit the disease to one other individual approximately every 3.33 days) and recovery rate $\gamma = 0.1$ (i.e. the average infectious period is 10 days). In this case, $R_0 = \frac{0.3}{0.1} = 3 > 1$. The number of susceptibles sharply declines as the disease spreads rapidly due to each infected individual transmitting the disease to more than one other person. The number of recovered rises steadily, since all individuals who get infected eventually recover. The number of infected individuals initially increases as more susceptible individuals are available to sustain transmission. Subsequently, the epidemic reaches its peak. Following this, the number of new cases declines and eventually ceases to change as fewer susceptible individuals are available to infect. The number of infections declines to zero over time since no infective individuals are left in the closed population.

On the other hand, in Figure 2(b), $\beta = 0.6$ and $\gamma = 0.8$, so that $R_0 = 0.75 < 1$; the disease rapidly disappears. These two plots demonstrate the impact of R_0 on disease dynamics.

In summary, when $R_0 > 1$, the disease initially spreads rapidly, peaks as the number of susceptible individuals decreases due to infection, and eventually dies out as there will be no infective individuals left in the closed population. Conversely, when $R_0 < 1$, the infection naturally diminishes over time as each infected individual spreads the disease to less than one other person.



(a) $R_0 > 1$ **(b)** $R_0 < 1$

Figure 2: Numerical solutions for deterministic SIR model using Euler's method for the parameters: (a) $\beta = 0.3$ and $\gamma = 0.1$ (for $R_0 > 1$) and (b) $\beta = 0.6$ and $\gamma = 0.8$ ($R_0 \leq 1$) with $N = 100$, $I_0 = 5$. The resulting plots illustrate the susceptible, infectious, and recovery populations by blue, red and green curves, respectively.

2.2. Limitations of deterministic SIR models

While the deterministic SIR model provides a valuable initial framework for understanding epidemic dynamics, it suffers from several limitations. These limitations highlight the need for more sophisticated modelling approaches, such as stochastic models, especially when dealing with real-world scenarios. Key limitations of the SIR model include:

Closed Population Assumption: The deterministic SIR model assumes a constant, closed population, neglecting the movement of individuals into or out of it. In reality, populations are often open, with migration and the introduction of external infections significantly impacting disease spread. For example, international travel played a crucial role in the rapid global spread of COVID-19.

Deterministic Nature: The model is deterministic, meaning that given initial conditions and parameters, the outcome is exactly predictable at any given time. This fails to account for the inherent randomness and variability in biological and social processes that influence disease transmission. Real-world epidemics exhibit unpredictable patterns due to chance events in individual contacts and disease progression.

Homogeneous Mixing: The assumption of homogeneous mixing, where every individual has an equal probability of contact with every other individual, is a simplification of complex social interaction patterns. Heterogeneities in contact structures can significantly affect disease spread.

Nonlinearity: The nonlinear nature of the differential equations can make analytical solutions difficult to obtain for more complex variations of the model.

Ignoring Stochastic Effects: Deterministic models are inadequate for describing the initial stages of an epidemic when the number of infected individuals is small, where stochas-

tic fluctuations can play a significant role in determining whether an outbreak occurs or fades out. They also fail to capture the possibility of disease extinction even when the conditions for spread ($R_0 > 1$) are met, due to random variations.

Inability to Explain Variability in Real-World Epidemics: Different regions or countries can experience significantly different epidemic patterns for the same disease even with seemingly similar conditions, a phenomenon that deterministic models struggle to explain due to their lack of inherent randomness.

These limitations of the deterministic SIR model underscore the necessity of employing stochastic models to provide a more realistic and comprehensive understanding of infectious disease dynamics, particularly in open populations influenced by external factors. It should be noted, however, that deterministic spatial models have been developed that allow movement of individuals between multiple populations (see for example Chen *et al.* (2020) and references therein). In this paper we focus only on the single population scenario.

3. Simple linear birth-death process in epidemiology

In a simple linear birth-death process applied to epidemiology, it is often assumed that the susceptible population is large enough to be considered infinite, resulting in an open population model where only the dynamics of the infected population are explicitly tracked. The rates of birth (infection) and death (recovery/death) are assumed to be linearly dependent on the current number of infected individuals n:

- **Birth rate (λ_n)**:
$$\lambda_n = n\lambda$$
where λ is the per-capita infection rate. This implies that the total rate of new infections is proportional to the number of currently infected individuals.

- **Death rate (μ_n)**:
$$\mu_n = n\mu$$
where μ is the per-capita recovery/death rate. This implies that the total rate of removal of infected individuals is proportional to the number of currently infected individuals.

When the number of infected individuals, n, reaches zero, the birth rate $\lambda_0 = 0$, indicating that state 0 is an absorbing state. Once the process reaches this state, the infection becomes extinct and the number of infected individuals remains at zero indefinitely. This process has been well studied by others (see Allen (2010); Bailey (1991); Kobayashi (2020)) and we give a brief summary of what is known along with example simulations in the remainder of this section.

3.1. Probability generating function and time-dependent distribution

The time-dependent probability mass function $P_n(t) = \mathbb{P}(I(t) = n \mid I(0) = I_0)$ gives the probability of having n infected individuals at time t given an initial number of

infections I_0. These state probabilities satisfy the forward Kolmogorov differential equations:

$$\frac{dP_n(t)}{dt} = \lambda(n-1)P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - (\lambda+\mu)nP_n(t), \quad n \geq 1. \tag{1}$$

When $n = 0$, $\lambda_0 = 0$ because $\lambda_n = n\lambda$ and we have:

$$\frac{dP_0(t)}{dt} = \mu P_1(t), \quad n = 0.$$

These are differential-difference equations and can be explicitly solved using the probability or moment-generating functions (PGF) approach. The PGF $G(z,t)$ is defined as:

$$G(z,t) = \sum_{n=0}^{\infty} z^n P_n(t).$$

For the simple linear birth-death process with $I(0) = I_0$, the PGF, $G(z,t)$, is given by (see Allen (2010); Bailey (1991); Kobayashi (2020)):

$$G_{\mathrm{BD}:I_0}(z,t) \equiv G(z,t) = \left(\frac{\frac{\mu(e^{at}-1)}{\lambda e^{at}-\mu} + \frac{1-\left(\mu(e^{at}-1)\right)}{\lambda e^{at}-\mu} - \frac{\lambda(e^{at}-1)}{\lambda e^{at}-\mu} z}{1 - \left(\frac{\lambda(e^{at}-1)}{(\lambda e^{at}-\mu)}\right) z} \right)^{I_0}, \tag{2}$$

where $a = \lambda - \mu$.

From this PGF, the time-dependent probability mass function $P_n(t)$ can be derived as:

$$P_n(t) = \sum_{j=0}^{\min(I_0,n)} \binom{I_0}{j} \binom{I_0+n-j-1}{I_0-1} \alpha(t)^{I_0-j} \beta(t)^{n-j} (1-\alpha(t)-\beta(t))^j$$

for $n \geq 1$, where

$$\alpha(t) = \frac{\mu(e^{at}-1)}{\lambda e^{at}-\mu}, \qquad \beta(t) = \frac{\lambda(e^{at}-1)}{\lambda e^{at}-\mu}.$$

The probability of extinction is:

$$P_0(t) = \alpha(t)^{I_0}.$$

3.2. Expected value and variance

$$E\left[I(t)\right] = \begin{cases} I_0 e^{at}, & \text{if } \lambda \neq \mu \\ I_0, & \text{if } \lambda = \mu. \end{cases} \tag{3}$$

The variance $\mathrm{Var}(I(t))$ is given by:

$$\mathrm{Var}(I(t)) = \begin{cases} I_0 \cdot \frac{\lambda+\mu}{\lambda-\mu} \cdot e^{(\lambda-\mu)t} \cdot \left(e^{(\lambda-\mu)t} - 1\right) & \text{if } \lambda \neq \mu \\ 2I_0\lambda t & \text{if } \lambda = \mu. \end{cases}$$

The variance demonstrates that the stochastic fluctuations around the expected value can be significant, especially over time.

3.3. Simulation of the birth-death process

To visualize the dynamic behaviour of the number of infected individuals over time within the stochastic framework, simulation of sample paths is crucial. This involves simultaneously computing the transition probabilities and the holding times in each state. The holding time in state n is exponentially distributed with rate $n(\lambda + \mu)$.

When a transition occurs, it is either a:

- **Birth (infection)** with probability $\lambda/(\lambda + \mu)$, or a
- **Death (recovery/death)** with probability $\mu/(\lambda + \mu)$.

Simulations reveal that individual sample paths can exhibit significant variability around the expected trajectory. Importantly, simulations demonstrate that even when $\lambda > \mu$ and the expected number of infected individuals grows exponentially, there is always a positive probability of extinction (reaching the absorbing state 0) in the stochastic model. See Figure 3.

This highlights a key difference from the deterministic model, where extinction is impossible if the birth rate exceeds the death rate.

3.4. Limitations of the linear birth-death model

While the linear birth-death model incorporates stochasticity, it also has limitations. The most significant limitation in the context of long-term epidemic dynamics is the absorbing state at zero. Once the number of infected individuals reaches zero, the disease becomes permanently extinct in the model, and no further infections can occur. This is unrealistic for many real-world diseases that can be reintroduced into a population through external sources or re-emergence. This limitation motivates the exploration of birth-death-immigration models, which account for the introduction of new infections from outside the population.

4. Birth-death immigration models

To overcome the limitation of the simple birth-death model regarding disease extinction, birth-death-immigration models incorporate the introduction of new infections from external sources, termed “immigration”. This section examines two types of birth-death-immigration models relevant to infectious disease dynamics in open populations: all-state immigration [Kobayashi (2020)] and zero-state immigration [Shiny and Viswanath (2022)].

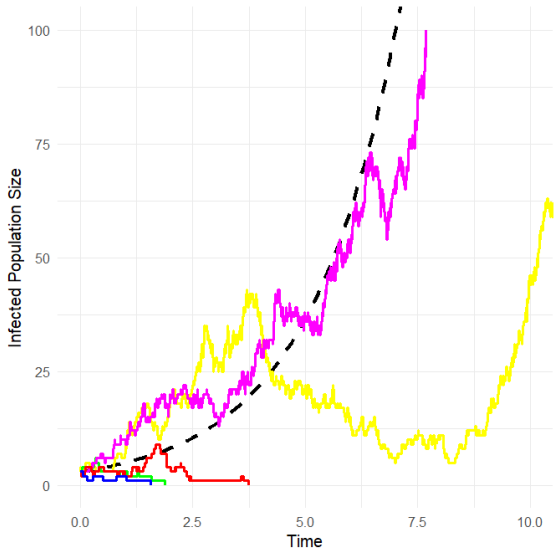


Figure 3: The stochastic mean (black dashed curve) and five sample paths of the birth and death model are plotted for $\lambda = 2$, $\mu = 1.5$, and $I_0 = 3$. Three curves (red, blue, green) from five sample paths quickly reach absorption while the remaining two curves (pink, yellow) continue.

4.1. Birth-death model with external infections via all-state immigration

The birth-death all-state immigration model extends the linear birth-death process by including a constant rate ν at which infected individuals enter the population from outside, regardless of the current number of infected individuals in the system. This model is characterized by three parameters:

- λ (Per-capita internal infection rate): As in the birth-death model, contributing to new infections within the population.
- μ (Per-capita recovery/death rate): Representing the removal of infected individuals.
- ν (External infection rate): A constant rate at which new infected individuals immigrate into the population per unit time.

The total infection rate when there are n infected individuals becomes $\lambda_n = n\lambda + \nu$, and the total recovery/death rate remains $\mu_n = n\mu$. This model allows for the persistence of the disease in the population, as new infections can continuously be introduced even when the number of internally generated infections is low.

4.1.1. Probability generating function and time-dependent distribution

Let $I_0 = I(0)$ denote the initial number of infected individuals in the system and $P_n(t)$ be the time-dependent probability mass function of the process $I(t)$:

$$P_n(t) = P(I(t) = n \mid I(0) = I_0), \quad n = 0, 1, 2, \ldots \quad \text{and} \quad t \geq 0,$$

with initial condition $P_n(0) = \delta_{n,I_0}$, where δ_{n,I_0} is the Kronecker delta function. The state probabilities, $P_n(t)$, in the birth-death immigration model satisfy the following Forward Kolmogorov equations:

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\nu P_0(t) + \mu P_1(t), \quad n = 0, \\
\frac{dP_n(t)}{dt} &= [(n-1)\lambda + \nu]P_{n-1}(t) - [n(\lambda+\mu)+\nu]P_n(t) + (n+1)\mu P_{n+1}(t), \quad n = 1, 2, 3, ...
\end{aligned} \tag{4}$$

with initial condition $I(0) = I_0$, i.e., $P_n(0) = \delta_{n,I_0}$ for $n \in \{0, 1, 2, 3, ...\}$.

The time-dependent PGF for the birth-death all-state immigration model with $I(0) = I_0$ is given by:

$$G(z,t) = G_{\text{BDI:0}}(z,t) \cdot G_{\text{BD:}I_0}(z,t),$$

where $G_{\text{BDI:0}}(z,t)$ is the PGF for the model with $I(0) = 0$, and $G_{\text{BD:}I_0}(z,t)$ is the PGF for the simple birth-death model with initial number of infections I_0 as in equation (2).

For the case with $I(0) = 0$, the PGF is:

$$G_{\text{BDI:0}}(z,t) = \left(\frac{\mu(e^{at}-1)}{\lambda e^{at} - \mu} + \frac{\nu(e^{at}-1)}{\lambda e^{at} - \mu} \cdot \frac{1-z}{\lambda - \mu z} \right)^{-\nu/\lambda}$$

where $a = \lambda - \mu$. This corresponds to a generalized negative binomial distribution.

The probability mass function for $I(0) = 0$ is:

$$P_n(t) = \binom{n+r-1}{n} (1-\beta(t))^r \beta(t)^n,$$

where $r = \nu/\lambda$ and $\beta(t) = \frac{\lambda(e^{at}-1)}{\lambda e^{at} - \mu}$.

4.1.2. Expected value and variance

The expected number of infected individuals for the birth-death all-state immigration model with $I(0) = I_0$ is:

$$\bar{I}(t) = \mathbb{E}[I(t)] = \begin{cases} I_0 e^{at} + \frac{\nu}{a}(e^{at}-1), & \text{if } a \neq 0 \\ I_0 + \nu t, & \text{if } a = 0 \end{cases}$$

where $a = \lambda - \mu$.

The variance is:

$$\text{Var}(I(t)) = \text{Var}_{\text{BD}}(I(t)) + \text{Var}_{\text{BDI:0}}(I(t))$$

$$= I_0 \cdot \frac{\lambda+\mu}{\lambda-\mu} e^{(\lambda-\mu)t} \left(e^{(\lambda-\mu)t} - 1 \right) + \frac{\nu}{\lambda-\mu} e^{(\lambda-\mu)t} \left(e^{(\lambda-\mu)t} - 1 \right), \quad \text{for } \lambda \neq \mu.$$

For $I(0) = 0$, the variance simplifies to:

$$\mathrm{Var}_{\mathrm{BDI:0}}(I(t)) = \frac{\nu}{\lambda - \mu} e^{(\lambda-\mu)t} \left(e^{(\lambda-\mu)t} - 1 \right).$$

A key feature of this model is that the variance of the number of infected individuals is often much higher than the mean, especially when $r = \nu/\lambda$ is small, leading to a high coefficient of variation (CV= $\frac{\sqrt{\mathrm{Var[I(t)]}}}{E[I(t)]} > 1$). This indicates significant over-dispersion and a greater degree of unpredictability in the epidemic patterns compared to what might be expected from a deterministic model or a process with lower variability. The generalized negative binomial distribution, which describes the probability distribution, has a long tail, implying a non-negligible probability of observing very large numbers of infected individuals. See Tables 1 ($r = \frac{\nu}{\lambda} = \frac{0.5}{0.6} \approx 0.83$) and 2 ($r = \frac{\nu}{\lambda} = \frac{0.1}{0.6} \approx 0.16$).

Table 1: The mean, variance and coefficient of variation for linear $BD:1$, $BDI:0$ and $BDI:1$, where per-capita infectious rate $\lambda = 0.6$, per-capita death/recovery rate $\mu = 0.4$, immigration rate $\nu = 0.5$, $a = \lambda - \mu = 0.2$.

	BD:I_0	BDI:0	BDI:I_0
$E[I(t)]$	54.5982	133.9954	188.5935
Var$[I(t)]$	14631.8	21679.71	36311.51
CV$[I(t)]$	2.2155	1.0988	1.0104

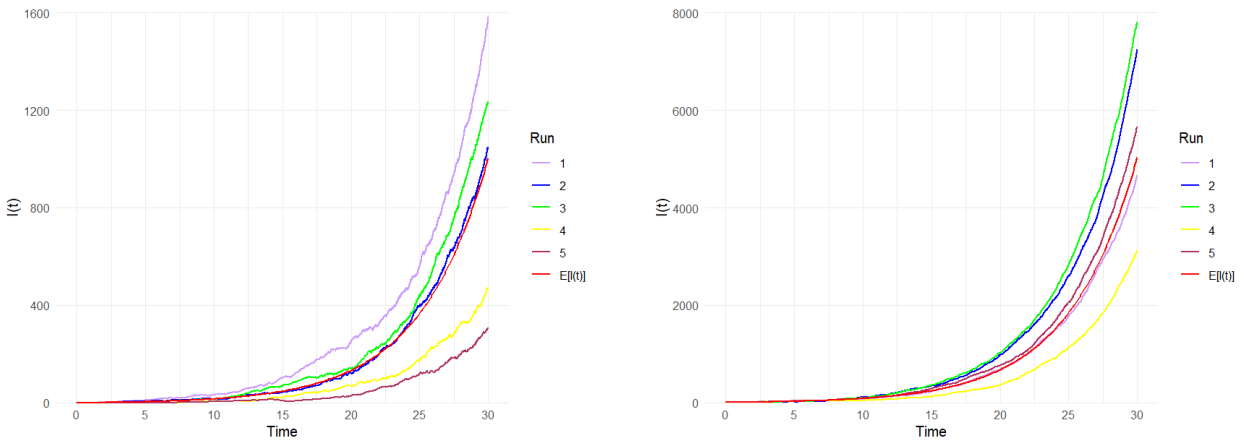
Table 2: The mean, variance, and CV for $BD:1$, $BDI:0$, and $BDI:1$ for $\lambda = 0.6$, $\mu = 0.4$, $\nu = 0.1$.

	BD:I_0	BDI:0	BDI:I_0
$E[I(t)]$	54.5982	26.780	81.3972
Var$[I(t)]$	14631.8	4335.942	18967.74
CV$[I(t)]$	2.2155	2.4571	1.6920

4.1.3. Simulation results and percentile curves

Simulations of the birth-death all-state immigration model, as illustrated in figure 4, show considerable variability between individual sample paths, even under the same parameter values and initial conditions. This reinforces the idea that chance can play a significant role in the actual progression of an epidemic. The long tail of the probability distribution implies that extreme scenarios, with much higher numbers of infections than the expected value, have a non-zero probability of occurring.

Percentile curves, derived from the probability distribution and simulation results (Figures 5 and 6), further highlight this variability. These curves show the range within which a certain percentage of the simulated epidemic trajectories are expected to fall over time. The wide spread between the percentile curves, especially at later times when exponential growth (if $\lambda > \mu$) becomes dominant, emphasizes the uncertainty in predicting the exact number of infections. The median of the distribution often deviates significantly from the expected value, further indicating the skewness and the limitations of relying solely on the mean for predictions.



(a) $I_0 = 0$ (b) $I_0 = 10$

Figure 4: Five sample paths for $I(t)$ versus t corresponding to the birth-death all-state immigration model. The expected number of infected individuals (red) is shown for (a) with $I_0 = 0$ and (b) with $I_0 = 10$. Parameters: $\lambda = 0.6$, $\mu = 0.4$, $\nu = 0.5$.

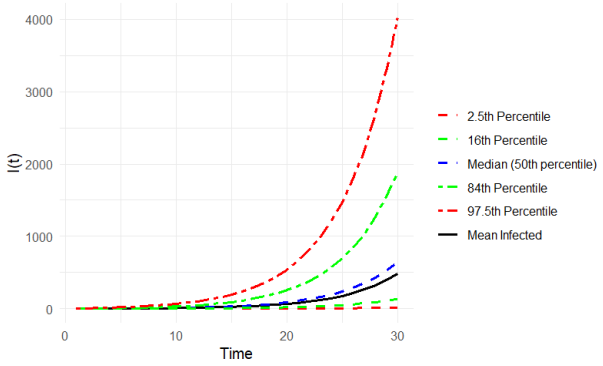


Figure 5: Percentile curves of $P_n(t)$ for $I(t)$ when $I_0 = 0$, $\lambda = 0.6$, $\mu = 0.4$, and $\nu = 0.5$ along with $\bar{I}(t)$.

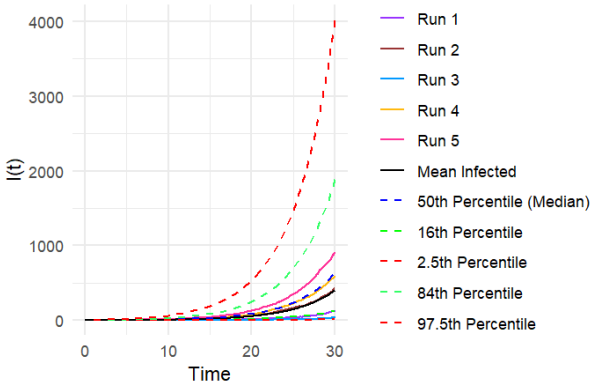


Figure 6: Percentile curves of $P_n(t)$ for simulated $I(t)$ when $I_0 = 0$, $\lambda = 0.6$, $\mu = 0.4$, and $\nu = 0.5$ along with $\bar{I}(t)$.

4.1.4. Stationary probability distribution

When $\lambda \leq \mu$ (i.e., $a \leq 0$), the birth-death all-state immigration model reaches a stationary probability distribution:

$$\pi_n = \lim_{t\to\infty} P_n(t).$$

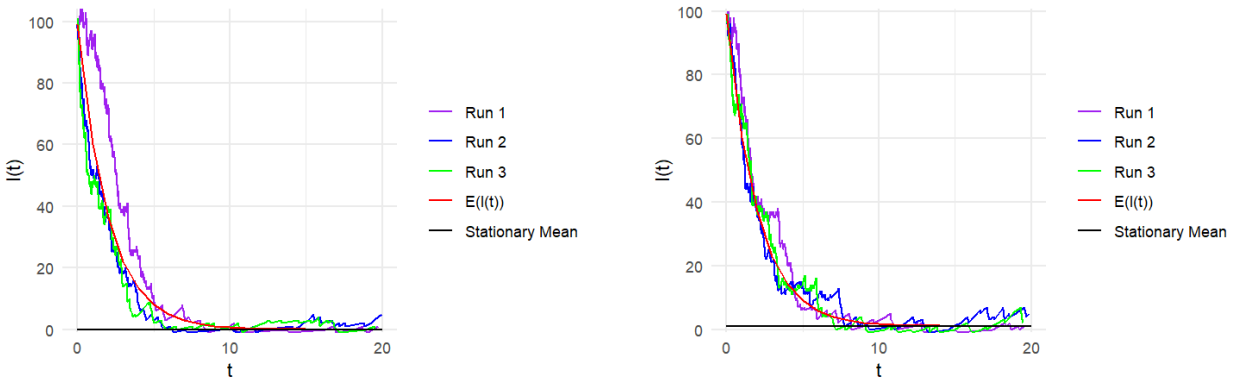
This distribution is also a generalized negative binomial distribution given by:

$$\pi_n = \binom{n+r-1}{n} \left(\frac{a}{\lambda}\right)^r \left(1 - \frac{a}{\lambda}\right)^n \tag{5}$$

where $r = \nu/\lambda$ and $a = \mu - \lambda$. The mean of this stationary distribution is:

$$\mathbb{E}[I] = \frac{\nu}{\mu - \lambda}, \quad \text{for } \lambda < \mu.$$

Simulations shown in Figure 7 illustrate how the number of infected individuals fluctuates around this stationary mean when $\lambda \leq \mu$.



(a) $\nu = 0.5$ (b) $\nu = 1$

Figure 7: Three sample paths of $I(t)$ versus time t for the birth-death all-state immigration model along with the expected value $E[I(t)] = \bar{I}(t)$ (red-line) when $I(0) = 100$, $\lambda = 0.5$, $\mu = 1$, and (a) $\nu = 0.5$, (b) $\nu = 1$; the respective mean values of the stationary distribution are 1, 2 (horizontal lines).

4.1.5. Associated processes of birth-death model with immigration

In this section, we study the deterministic mean behaviour of the associated processes in the Birth-Death model with immigration. In addition to the process describing the number of infected individuals, $I(t)$, $t \geq 0$, we introduce the following processes:

1. The process $\{A(t), t \geq 0\}$ denotes the cumulative number of infectious individuals entering the population from outside by time t. We assume that $A(t)$ is a Poisson process with rate ν (individuals per unit time).

2. The process $\{B(t), t \geq 0\}$ denotes the cumulative number of internally generated infections by time t. Secondary infections are assumed to occur at rate λ (persons/unit time/infectious person).

3. The process $\{R(t), t \geq 0\}$ denotes the cumulative number of recovered or deceased individuals by time t. Such departures are assumed to occur at rate μ (persons/unit time/infected person).

Let $I_0 = I(0)$ represent the initial number of infected individuals. Then, the number of infected individuals in the population at time t is given by

$$I(t) = I(0) + A(t) + B(t) - R(t). \tag{6}$$

The external arrival process $(A(t), t \geq 0)$ is a Poisson process with rate ν and is therefore well characterized. However, the processes $(B(t), t \geq 0)$ and $(R(t), t \geq 0)$ are more difficult

to analyze. Nevertheless, expressions for their expected values can be obtained from the mean behavior of the infection process $(I(t), t \geq 0)$.

Using the probability generating function associated with $(I(t), t \geq 0)$, the ordinary differential equation governing the mean number of infected individuals, $\bar{I}(t) = E[I(t)]$, is obtained as

$$\frac{d\bar{I}(t)}{dt} = (\lambda - \mu)\bar{I}(t) + \nu. \tag{7}$$

Solving the above differential equation yields

$$\bar{I}(t) = E[I(t)] = I_0 e^{at} + \frac{\nu}{a}(e^{at} - 1), \quad I_0 = I(0) \quad \text{and} \quad a = \lambda - \mu, \tag{8}$$

where $I_0 = I(0)$.

Since $A(t)$ is a Poisson process with rate ν, the expected number of external infectious arrivals by time t is

$$E[A(t)] = \bar{A}(t) = \nu t, \quad t \geq 0. \tag{9}$$

Because infected individuals generate secondary infections at rate λ, the expected cumulative number of internal infections, $\bar{B}(t)$, satisfies

$$\frac{d\bar{B}(t)}{dt} = \lambda \bar{I}(t). \tag{10}$$

Integrating the above equation and using the expression for $\bar{I}(t)$ gives:

$$\begin{aligned} E[B(t)] = \bar{B}(t) &= \lambda \int_0^t \bar{I}(u)\, du, \\ &= \frac{\lambda I_0(e^{at} - 1)}{a} + \frac{\lambda\nu(e^{at} - 1)}{a^2} - \frac{\lambda\nu t}{a}, \\ &= \frac{\lambda}{a}(\bar{I}(t) - I_0 - \bar{A}(t)). \end{aligned} \tag{11}$$

Similarly, since the infected population recovers/dies at the rate of μ, the expected number of recovered or deceased individuals $\bar{R}(t)$, satisfies

$$\frac{d\bar{R}(t)}{dt} = \mu \bar{I}(t). \tag{12}$$

Integrating both sides and using Equation (8), we obtain the solution for the ODE (12):

$$\begin{aligned} E[R(t)] = \bar{R}(t) &= \mu \int_0^t \bar{I}(u)\, du, \\ &= \frac{\lambda I_0(e^{at} - 1)}{a} + \frac{\lambda\nu(e^{at} - 1)}{a^2} - \frac{\lambda\nu t}{a}, \\ &= \frac{\mu}{a}(\bar{I}(t) - I_0 - \bar{A}(t)). \end{aligned} \tag{13}$$

Although explicit expressions for the expected values of the associated processes can be obtained, a detailed characterization of the stochastic behaviour of the processes $(B(t), t \geq 0)$ and $(R(t), t \geq 0)$ remains an open problem and is currently under investigation. In particular, the process $(R(t), t \geq 0)$ resembles the departure process of a queueing system, and such processes are generally difficult to characterize analytically.

4.2. Birth-death model with immigration only at zero (zero-state immigration)

The birth-and-death zero-state immigration model presents an alternative scenario in which external infections (immigration at a rate ν) are introduced only when the number of infected individuals in the system reaches zero. This can model situations where strict control measures, such as travel bans, are implemented as soon as infections are detected and are only lifted when the infection is eradicated.

The infinitesimal transition probabilities for this model differ from the all-state immigration model, particularly at state 0:

$$\begin{aligned} P[I(t+\Delta t) = 1 \mid I(t) = 0] &= \nu \Delta t + o(\Delta t) \\ P[I(t+\Delta t) = n+1 \mid I(t) = n] &= n\lambda \Delta t + o(\Delta t), \quad n \geq 1 \\ P[I(t+\Delta t) = n-1 \mid I(t) = n] &= n\mu \Delta t + o(\Delta t), \quad n \geq 1. \end{aligned}$$

The differential-difference equations for $P_n(t)$ in this model are more complex. $P_n(t) = P(I(t) = n \mid I(0) = I_0)$ satisfies the following differential-difference equations:

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\nu P_0(t) + \mu P_1(t), \quad n = 0, \\ \frac{dP_1(t)}{dt} &= \nu P_0(t) - (\lambda + \mu) P_1(t) + 2\lambda P_2(t), \quad n = 1, \\ \frac{dP_n(t)}{dt} &= (n-1)\lambda P_{n-1}(t) - n(\lambda + \mu) P_n(t) + (n+1)\mu P_{n+1}(t), \quad n \geq 2. \end{aligned} \tag{14}$$

Assume that $I(0) = 1$ and hence $P_1(0) = 1$. Using the probability-generating function approach, we can transform the Kolmogorov forward equations into a partial differential equation:

$$\frac{\partial G(z,t)}{\partial t} = (z-1)\left[(\lambda z - \mu)\frac{\partial G(z,t)}{\partial z} + \nu P_0(t)\right], \tag{15}$$

with initial condition:

$$G(z,0) = z, \quad G(1,t) = 1.$$

Finding an exact solution to this equation is not straightforward due to the presence of the $P_0(t)$ term. However, one can find,

$$G(z,t) = \int_0^{\zeta} \nu \frac{\phi(u)}{1-\phi(u)} \left(1 - \frac{\mu}{\lambda}\right) p_0(u) du + \zeta.$$

and

$$P_0(t) = \int_0^t \nu \frac{\phi(u)}{1-\phi(u)} \left(1 - \frac{\mu}{\lambda}\right) P_0(u)du + \frac{1}{1 - \frac{\lambda}{\mu}\exp((\lambda-\mu)t)} - \frac{\mu}{\lambda}. \tag{16}$$

where

$$\phi(\zeta) = \exp((\mu-\lambda)\zeta)\left(\frac{\zeta-1}{\zeta-\frac{\mu}{\lambda}}\right).$$

This can be used to find the expected value and the variance of the system.

Deriving the mean and variance using the probability generating function is straightforward, and they are given by:

$$E[I(t)] = \exp((\lambda-\mu)t)\left(\int_0^t \nu \exp((\mu-\lambda)u)P_0(u)du + 1\right);$$

$$\mathrm{Var}[I(t)] = \exp(2(\lambda-\mu)t)\left(\int_0^t (\nu P_0(u)(1-2E(u)) + (\lambda+\mu)E(u))\exp(2(\mu-\lambda)u)du\right).$$

Both expressions involve $P_0(t)$ and cannot be computed exactly; one has to use numerical techniques for Volterra-type integral equations to calculate $P_0(t)$ (see Shiny and Viswanath (2022)). However, a stationary probability distribution exists for this model if $\lambda < \mu$, given by:

$$\pi_0 = \frac{1}{1 - \frac{\nu}{\mu}\ln\left|1 - \frac{\lambda}{\mu}\right|} \tag{17}$$

$$\pi_n = \frac{\nu}{n\mu}\left(\frac{\lambda}{\mu}\right)^{n-1}\pi_0, \quad n = 1, 2, 3, \ldots \tag{18}$$

This model, while also incorporating external infections, captures a different dynamic where the disease can experience periods of extinction followed by reintroduction, making it relevant for understanding the persistence and re-emergence of certain infectious diseases.

5. Model extension: Time-nonhomogeneous parameters

Real-world epidemic scenarios are often characterized by changes in the rates of infection, recovery, and external influences over time due to factors such as public health interventions, changes in population behaviour, and variations in the influx of infected individuals. To better reflect these dynamic conditions, the birth-death all-state immigration model can be extended to incorporate time-nonhomogeneous parameters: $\lambda(t)$, $\mu(t)$, and $\nu(t)$.

For this non-homogeneous model, the expected number of infected individuals at time t, given $I(0) = I_0$, is given by Kobayashi's formula:

$$\bar{I}(t) = I_0 \cdot e^{s(t)} + e^{s(t)} \cdot N(t),$$

where

$$s(t) = \int_0^t [\lambda(u) - \mu(u)]\, du, \quad N(t) = \int_0^t \nu(u) \cdot e^{-s(u)}\, du.$$

This extension allows for the analysis of how time-varying rates impact the expected epidemic trajectory. This section next explores three specific scenarios under this framework with constant recovery rate μ: varying λ with constant ν; varying ν with constant λ; varying λ and ν.

5.1. Analysis with time-nonhomogeneous infection rate $\lambda(t)$

This scenario considers a situation where the per-capita infection rate $\lambda(t)$ changes over time while the recovery rate μ and immigration rate ν remain constant. This is particularly relevant for modeling the impact of interventions aimed at reducing transmission, such as social distancing measures or lock downs.

This section reproduces an example from Kobayashi, where $\lambda(t)$ is modeled as a decreasing function over a period of time d, representing the time it takes for the public to fully respond to and implement government regulations aimed at reducing contact rates:

$$\lambda(t) = \begin{cases} \lambda_0, & t < t_1 \\ \lambda_1 + (\lambda_0 - \lambda_1) \cdot \frac{1-\cos\left(\pi \cdot \frac{t-t_1}{d}\right)}{2}, & t_1 \le t \le t_1 + d \\ \lambda_1, & t > t_1 + d, \end{cases}$$

where t_1 is the time at which the intervention is announced, λ_0 is the initial infection rate, and λ_1 is the reduced infection rate. See Figure 8.

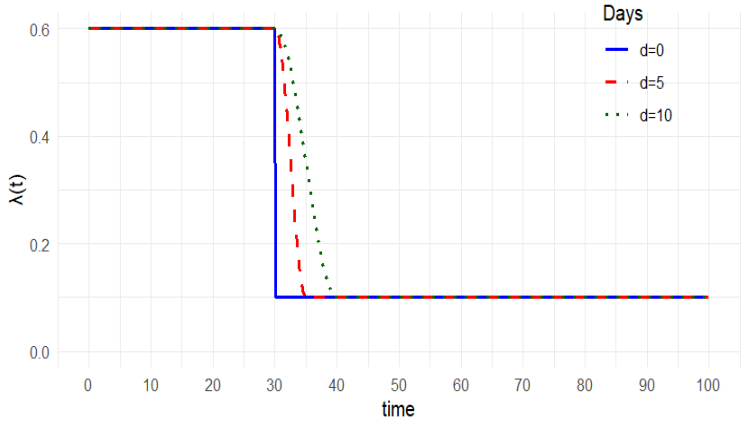


Figure 8: A plot of $\lambda(t)$ from time $t = 0$ to $t = 100$ and d takes values of 0, 5, and 10.

Analysis of the expected number of infected individuals under such a time-varying $\lambda(t)$ demonstrates the significant impact of the delay d in the public's response. See Figure 9. A longer delay leads to a more rapid initial spread and a higher peak in the number of infections. This underscores the critical importance of prompt and effective implementation of control

measures to mitigate the impact of an outbreak. To better reflect realistic scenarios, our model could be extended to implement a finite duration $D > d$ for the control measures, for example by resetting $\lambda(t) = \lambda_0$ for $t \geq t_1 + D$. We expect that infections would rise again after the control period is over, however, perhaps not to the original level, reflecting the impact of the initial controls. This is left for future work.

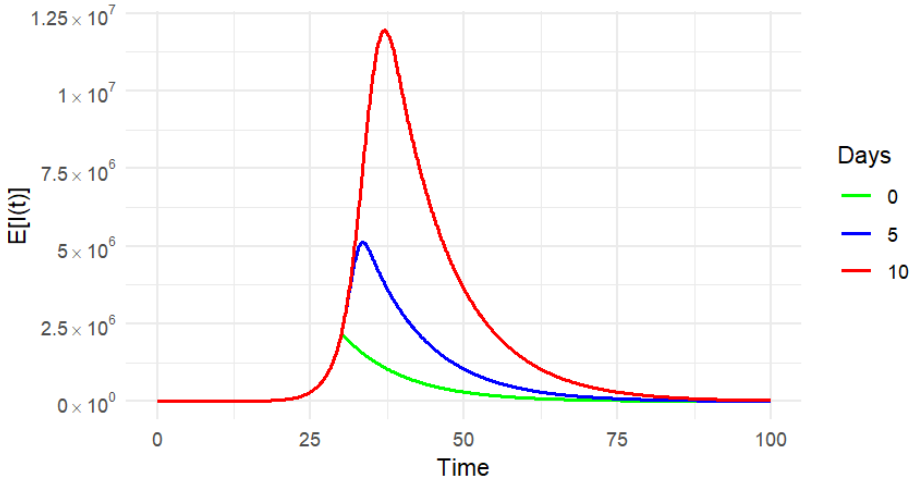


Figure 9: The expected number of infected individuals versus time under the conditions of $\lambda(t)$, $I_0 = 1$, $\nu = 5$, $\mu = 0.2$. The results for different d values are represented by three colours: blue for $d = 0$, green for $d = 5$, and red for $d = 10$.

5.2. Analysis with time-nonhomogeneous immigration rate $\nu(t)$

This scenario examines the effect of a time-varying external infection rate $\nu(t)$ while keeping λ and μ constant. This is relevant for understanding the impact of changing travel patterns, border controls, or the emergence of new external sources of infection.

This section considers $\nu(t)$ modeled by a step function and a piecewise function representing a gradual decline in immigration (see Figure 10):

$$\nu(t) = \begin{cases} \nu_0, & 0 \leq t < t_1 \\ \nu_1, & t \geq t_1 \end{cases} \quad \text{and} \quad \nu(t) = \begin{cases} \nu_0, & t < t_1 \\ \nu_0 + (\nu_1 - \nu_0) \cdot \frac{1-\cos\left(\pi \cdot \frac{t-t_1}{d}\right)}{2}, & t_1 \leq t \leq t_1 + d \\ \nu_1, & t > t_1 + d. \end{cases}$$

Analysis of the expected number of infected individuals under these scenarios highlights how changes in the rate of external infections can influence the overall epidemic trajectory. See Figure 11. A sudden reduction in $\nu(t)$ leads to a flattening of the curve and a lower equilibrium level of infection (if $\lambda \leq \mu$). Similar to the case with $\lambda(t)$, delays in implementing measures to reduce external infections have a detrimental effect on controlling the spread. Similarly, our model could be extended to implement a finite duration $D > d$ for travel restrictions, for example by resetting $\nu(t) = \nu_0$ for $t \geq t_1 + D$. Since external infections will rise after the restriction period is over, we also expect the number of infections to rise. This is left for future work.

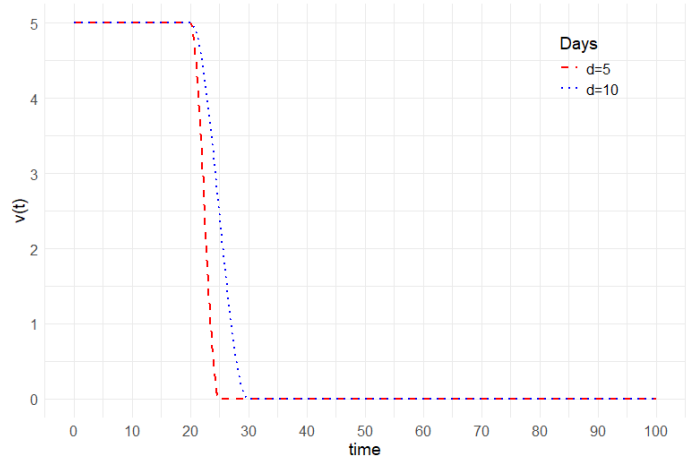


Figure 10: The function $\nu(t)$ transitions from $\nu_0 = 5$ to $\nu_1 = 0.001$ over the interval from $t_1 = 20$ to $t_1 + d = 20 + d$, where d takes values of 5 and 10. The results for different values of d are represented in three colors: red for $d = 5$, and blue for $d = 10$. Note that for $d = 0$, $\nu(t)$ is a step function which drops from $\nu_0 = 5$ to $\nu_1 = 0.001$ at $t = 20$.

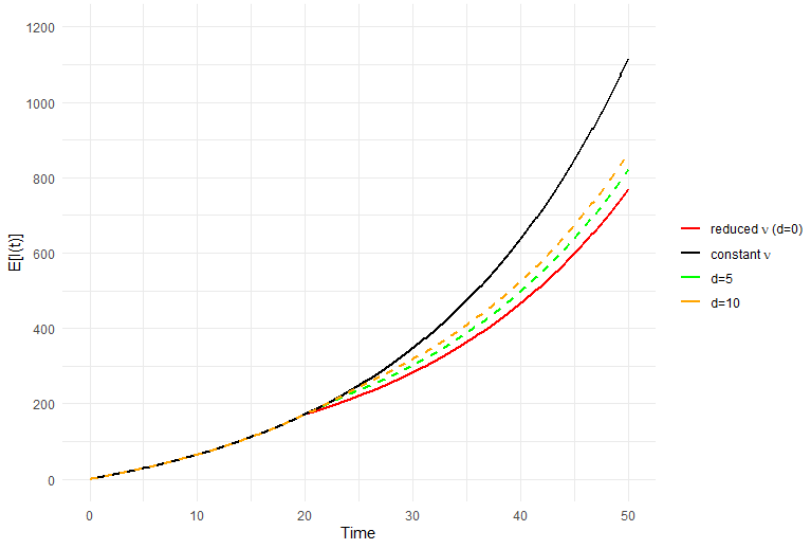


Figure 11: The expected number of infected individuals versus time in the system under the conditions of $\nu(t)$, $I_0 = 0$, $\lambda = 0.65$, $\mu = 0.6$. The results for different d values are represented in three colors: solid red for $d = 0$, dashed green for $d = 5$, and dashed orange for $d = 10$. The solid black curve represents the case where the transmission rate ν is constant at 5 for the entire time.

5.3. Analysis with time-nonhomogeneous $\lambda(t)$ and $\nu(t)$

Finally, in this section, we investigate the scenario where both the per-capita infection rate $\lambda(t)$ and the external infection rate $\nu(t)$ vary over time, while the recovery rate μ remains constant. This provides a more comprehensive model that accounts for the simultaneous effects of internal transmission control measures and changes in external exposures.

Using the same functional forms for $\lambda(t)$ and $\nu(t)$ as in the previous analyses, the expected number of infected individuals is calculated using the general formula for $\bar{I}(t)$. See Figure 12. Comparison of the results with the case where only $\lambda(t)$ is time-varying (Figure 9) suggests that while altering both rates simultaneously has an impact, reducing the internal infection rate $\lambda(t)$ has a more substantial effect on controlling the expected number of infections. This implies that focusing on measures that reduce transmission within the population may be more effective than solely focusing on reducing external introductions.

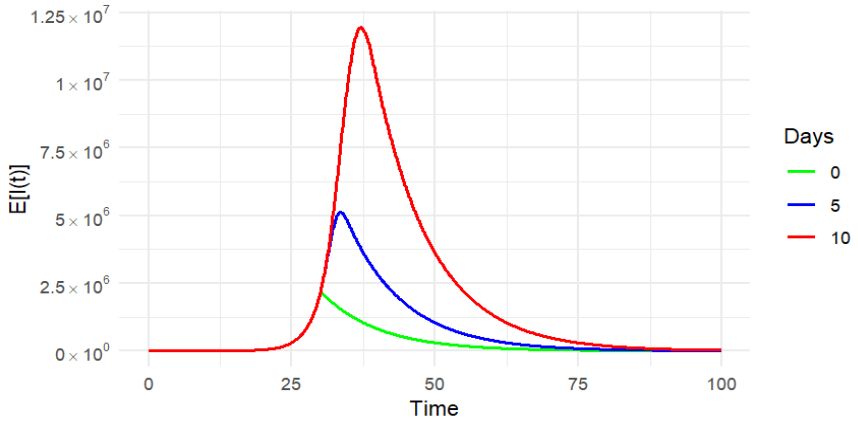


Figure 12: The expected number of infected individuals versus time under the conditions of $I_0 = 0$, $\lambda(t)$, $\nu(t)$, $\mu = 0.2$. The results for different d values are represented in three colors: blue for $d = 0$, green for $d = 5$, and red for $d = 10$.

6. Discussion and conclusion

This paper has explored the application of stochastic models, particularly birth-death and birth-death-immigration processes, to understand the dynamics of infectious disease spread in an open population, emphasizing the role of external infections. The limitations of the foundational deterministic SIR model, including its closed population assumption and inability to capture inherent randomness, necessitate the use of stochastic frameworks for more realistic epidemic modelling. Note that we focus on a single open population and assume it is so large that the number of susceptible individuals can be assumed to be unbounded. We do not consider multiple populations where migration of susceptible individuals could have a significant impact on the epidemic dynamics. Stochastic network models are expected to be suitable for modelling this scenario but that is beyond the scope of the models presented here.

The simple linear birth-death model, while incorporating stochasticity, suffers from the limitation of an absorbing state at disease extinction, which does not reflect the possibility of re-introduction in open populations. The birth-death all-state immigration model addresses this by allowing for a continuous influx of external infections, leading to a probability distribution of the number of infected individuals that follows a generalized negative binomial distribution with a long tail and high variability. This inherent variability, demonstrated through simulations and percentile curves, highlights the unpredictable nature of epidemic patterns and can explain why real-world outbreaks often deviate significantly from

deterministic predictions.

Introducing time-nonhomogeneous parameters into the birth-death all-state immigration model provides a powerful tool for analyzing the impact of changing conditions and interventions on epidemic dynamics. The analysis of varying internal infection rates $\lambda(t)$ underscores the crucial role of prompt and effective public health measures in mitigating outbreaks. Similarly, changes in the external infection rate $\nu(t)$, representing factors like travel restrictions, can significantly influence the epidemic trajectory. Notably, the research suggests that reducing the internal infection rate has a more profound impact on controlling the expected number of infections compared to solely focusing on reducing external introductions.

Much work remains to be done regarding applying birth-death-immigration models to epidemics. **Parameter Estimation:** Develop and apply statistical methods, such as maximum likelihood estimation or the Expectation-Maximization (EM) algorithm, to estimate the parameters of the birth-death all-state immigration stochastic model from real-world epidemiological datasets. This would allow for data-driven predictions of future epidemic behaviour. **Analysis of Cumulative Processes:** Conduct more detailed theoretical studies of the cumulative internal infection process $B(t)$ and the cumulative recovery/death process $R(t)$ within the birth-death-immigration framework to gain further insights into the overall dynamics of the epidemic. **Evaluation of Public Health Interventions:** Extend the time-nonhomogeneous models to explicitly incorporate and evaluate the impact of various public health techniques, such as quarantine strategies and different vaccination approaches, on disease spread and control. **Incorporating Heterogeneity:** Develop stochastic models that account for heterogeneity in population mixing, susceptibility, and infectivity to provide even more realistic representations of disease transmission dynamics. **Application to Specific Diseases:** Apply the developed stochastic models with external infection to analyze the dynamics of specific infectious diseases, utilizing real-world data to validate the models and generate disease-specific predictions.

In conclusion, this paper emphasizes the importance of stochastic modelling, particularly birth-death-immigration models with the capacity to incorporate time-varying parameters, for gaining a deeper understanding of infectious disease dynamics in an interconnected world where external infections are a significant factor. These models offer a more realistic framework for predicting epidemic behavior, assessing the impact of interventions, and informing public health strategies aimed at controlling and mitigating the spread of infectious diseases.

Acknowledgements

CS acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2020-06339] and of the Mathematics for Public Health (MfPH) program. HM acknowledges support from University of Saskatchewan, Department of Mathematics and Statistics, Graduate Teaching Fellowships. The authors also acknowledge helpful comments from an anonymous referee.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Allen, L. J. (2008). An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer.

Allen, L. J. (2010). *An Introduction to Stochastic Processes with Applications to Biology*. CRC press.

Allen, L. J., Brauer, F., Van den Driessche, P., and Wu, J. (2008). *Mathematical Epidemiology*, volume 1945. Springer.

Allen, L. J., Jang, S. R., and Roeger, L.-I. (2017). Predicting population extinction or disease outbreaks with stochastic models. *Letters in Biomathematics*, **4**, 1–22.

Andersson, H. and Britton, T. (2012). *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151. Springer Science & Business Media.

Bailey, N. T. (1991). *The Elements of Stochastic Processes with Applications to the Natural Sciences*, volume 25. John Wiley & Sons.

Brauer, F. (2017). Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, **2**, 113–127.

Chen, S., Owolabi, Y., Li, A., Lo, E., Robinson, P., Janies, D., Lee, C., and Dulin, K. M. (2020). Patch dynamics modeling framework from pathogens' perspective: Unified and standardized approach for complicated epidemic systems. *PLoS ONE*, **15**, e0238186.

Kendall, D. G. (1948a). On some modes of population growth leading to RA Fisher's logarithmic series distribution. *Biometrika*, **35**, 6–15.

Kendall, D. G. (1948b). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, **19**, 1–15.

Kobayashi, H. (2020). Stochastic modeling of an infectious disease, part I: Understand the negative binomial distribution and predict an epidemic more reliably. *eprint*, **arXiv:2006.01586 [q-bio.PE]**.

Kobayashi, H. (2021a). Stochastic modeling of an infectious disease part II: Simulation experiments and verification of the analysis. *eprint*, **arXiv:22101.11394 [q-bio.PE]**.

Kobayashi, H. (2021b). Stochastic modeling of an infectious disease part III-A: Analysis of time-nonhomogeneous models. *eprint*, **arXiv:2101.09109 [stat.ME]**.

Kobayashi, H. (2021c). Stochastic modeling of an infectious disease, part III-B: Analysis of the time-nonhomogeneous BDI process and simulation experiments of both BD and BDI processes. *eprint*, **arXiv:2104.00529 [math.PR]**.

Kretzschmar, M. and Wallinga, J. (2010). Mathematical models in infectious disease epidemiology. In Krämer, A., Kretzschmar, M., and Krickeberg, K., editors, *Modern Infectious Disease Epidemiology: Concepts, Methods, Mathematical Models, and Public Health*, pages 209–221. Springer New York, New York, NY.

Macdonald, G. (1952). The analysis of equilibrium in malaria. *Tropical Diseases Bulletin*, **49**, 813–829.

Madushani U.K.A., H. (2025). Stochastic modelling of infectious disease dynamics in open populations with external infections. Master's thesis, University of Saskatchewan.

Shiny, K. and Viswanath, N. C. (2022). Study of birth-death processes with immigration. *Croatian Operational Research Review*, **13**, 49–63.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 37-62

Professor VK Gupta's Influence on Design of Experiments and Sample Surveys

Rajender Parsad

Indian Council of Agricultural Research, New Delhi

Received: 26 April 2026; Revised: 28 July 2026; Accepted: 05 August 2026

Abstract

Professor V. K. Gupta has made notable contributions in Designing of Experiments that includes obtaining efficient and robust designs for various experimental settings *viz.* block designs; row-column designs; block designs with nested factors; designs for making test treatments-control comparisons; designs for diallel crosses; designs for bioassays; crop sequence experiments; interference effects; correlated error structures; fractional factorial plans, supersaturated designs; computer aided search of efficient designs for many experimental situations and for many practical ranges of parameters; shape and size of plots and blocks, diagnostics in designed field experiments; intercropping experiments; experiments with split application of inputs, *etc.* He has also made use of the combinatorics in sample surveys and because of this he has made several original contributions to the theory of survey sampling, particularly variance estimation from large scale complex survey data, unequal probability sampling, controlled selection and balanced and distance balanced sampling plans. He has also worked on calibration approach of estimation of parameters. In this manuscript, a valiant attempt shall be made to highlight briefly few prominent results from the above.

Key words: Design of experiments; Sample surveys; Optimal designs; Robust designs.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Professor V. K. Gupta (VKG) was born in Bombay, India, on March 17, 1951. He joined Indian Agricultural Statistics Research Institute (IASRI) [the then Institute of Agricultural Research Statistics (IARS)] in 1970 as M.Sc. student in Agricultural Statistics followed by P.G. Diploma in 1974. He started his academic journey during May 1974 from Chaudhary Charan Singh Haryana Agricultural University (CCS HAU), Hisar [the then Haryana Agricultural University (HAU)] and was there till December 1974. He joined ICAR-IASRI in 1975 and superannuated in March 2016 as National Professor at the age of 65 years. Now he has completed 55 years of his Journey in the discipline of Statistics.

Corresponding Author: Rajender Parsad
Email: rajender1066@yahoo.co.in

Describing the work done by him is showing lamp to the bright sun. But still, I am tempted to take this chance.

2. Academic qualifications

VKG got his school education at Delhi United Christian School, Delhi, and completed his senior Secondary in 1967. After completing B.Sc. from Ramjas College, University of Delhi in 1970, he joined IASRI (then IARS) as M.Sc. student in 1970 and completed his M.Sc. in 1972. During his M.Sc. he worked on **Study on efficiency of systematic sampling in estimating lactation yield of a cow** under the guidance of Sh. M. Rajagopalan. It was followed by P.G. Diploma in Agricultural Statistics in 1974 during which he worked on **Some empirical studies on estimation procedures in sample surveys on minor crops** under the guidance of Dr. A.H. Manwani. He completed his Ph.D. (Agricultural Statistics), as in-service student and submitted his Ph.D. thesis on the topic **Use of combinatorial properties of block designs in unequal probability sampling** in less than 2 years under the able guidance of Professor A. K. Nigam.

3. Professional career

VKG began his professional career in 1974 at HAU Hisar (now known as CCS HAU, Hisar) as Assistant Statistician. He then joined Institute of Agricultural Research Statistics (IARS) [now known as ICAR-IASRI] as Statistical Investigator in 1975. In 1976, Agricultural Research Service (ARS) was initiated by Indian Council of Agricultural Research and after passing the ARS Examination, he joined IASRI as Scientist S-1 in the first batch of ARS in August 1976. In 1982, he was promoted to Scientist S-2, which was re-designated as Senior Scientist in 1995. He coordinated the teaching activities of discipline of Agricultural Statistics with PG School, Indian Agricultural Research Institute (IARI) as Professor (Agricultural Statistics). Based on his significant research contributions, he was awarded the position of ICAR National Fellow in 1995. Perhaps he was the first ICAR National Fellow in the discipline of Agricultural Statistics (1995-2000). After 2000, by now four other Agricultural Statisticians (Rajender Parsad, Hukum Chandra, Ranjit Kumar Paul and Eldho Varghese) have received ICAR National Fellow Position. He has also contributed significantly to research management as Head, Division of Design of Experiments during 2000-2003 and Joint Director during 2003-2006 at IASRI (now known as ICAR-IASRI). He was invited by FAO as Consultant in 1994 in Beijing to design experiments for balanced fertilizers, a topic of global concern at present. He brought laurels and recognitions to the discipline of Agricultural Statistics in 2006 when he was awarded the position of ICAR National Professor for the first time to any Agricultural Statistician. He nurtured the school of Design of Experiments as ICAR National Professor, the position from which he superannuated on March 16, 2016.

4. Alma mater

ICAR-IASRI, New Delhi is a premier Institute of relevance in Statistical Sciences (Statistics, Computer Applications and Bioinformatics) and their judicious fusion in agricultural sciences for enriching quality agricultural research and informed policy decision making. It is in 10^{th} Decade in the service of Nation [(a) 96 years in Agricultural Statistics: 1930 Statistical section in then ICAR; (b) 61 years in Computer Applications: 1965 $\sim$ Installa-

tion of ICAR's first computing facility-IBM-1620 Data Processing System: All Statistical Analysis of data from National Agricultural Research and Education System (NARES) was done at ICAR-IASRI; and (c) 16 Years in Agricultural Bioinformatics: 2010 Establishment of National Agricultural Bioinformatics Grid].

ICAR-IASRI started its journey as a Statistical Section in 1930 in then Imperial Council of Agricultural Research and has grown to a premier institute of relevance to conduct research, education and training in the field of Statistical Sciences (Statistics, Computer Applications, and Bioinformatics). The Institute is mainly responsible for conducting research in Agricultural Statistics and Informatics to bridge the gaps in the existing knowledge. The Institute is using the power of Statistics, as a science, blended with Informatics and their judicious fusion in agricultural sciences for enhancing quality agricultural research, to meet the challenges of agricultural research in newer emerging areas and evidence-based policy decision making. The Institute also conducts M.Sc. and Ph.D. degree programmes in Agricultural Statistics, Computer Applications and Bioinformatics in collaboration with the Graduate School, ICAR-IARI, New Delhi. The Institute organizes customized and sponsored training courses in Agricultural Statistics and Informatics at National and International level to be a leading Centre of excellence in Human Resource Development. The Institute provides advisory and consultancy services for strengthening the National Agricultural Research and Education System (NARES) and undertakes sponsored research and consultancy for National and International organizations. The methodological support is provided in strengthening National Agricultural Statistics System (NASS). The Institute has also been playing a lead role in development of robust Agricultural Knowledge Management Systems and artificial intelligence-based applications for NARES. The present mandate of the ICAR-IASRI and major activities are

Mandate	Major Activities
• Research, education and training in agricultural statistics, computer applications in agriculture and agricultural bioinformatics. • Advisory / consultancy services / methodological support / computational solutions to NARES/NASS.	• Developing Statistical methodologies / techniques to support agricultural research and development. • Developing, supporting and maintaining Portals / Information-Management-Systems / Mobile Apps *etc.* and digital solutions for agricultural research and development. • Accelerating agricultural OMICS research in India through Network Project on Agricultural Bioinformatics and Computational Biology. • Capacity building in terms of M.Sc. and Ph.D. students in Statistical Sciences and knowledge updating of research fraternity through customized training programmes.

Stakeholders of ICAR-IASRI include Students, Scientists, Government Officials, Farmers. Throughout his professional career, VKG has been at his alma mater for 45+ years (more than half of 86 years journey of IASRI till 2016).

5. Awards and honours

VKG's monumental contributions have brought him many laurels, and he has been decorated with several awards like IARI Best Teacher Award 1999; First D. N. Lal Memorial Lecture award 2000 from Indian Society of Agricultural Statistics (ISAS); Prof. P. V. Sukhatme Gold Medal Award 2005 from ISAS; "Sankhyiki Bhusan, 2010 from ISAS", *etc.* He is an elected member of the International Statistical Institute; Elected Fellow of the National Academy of Agricultural Sciences (NAAS) and Fellow of the ISAS. He has been the Chair Editor, (i) Journal of the Indian Society of Agricultural Statistics during 2005-2012; and (ii) Statistics and Applications 2020-2025 and now Editor-in-Chief till date; Associate Editor of Journal of Statistical Planning and Inference; at present he is an Associate Editor, Journal of Statistical Theory and Practice and Agricultural Research. He is associated with many Professional Societies. He has been Sessional President, 61^{st} Annual Conference, ISAS, 2007; Vice-President of Indian Society of Agricultural Statistics; Vice-President of Forum for Interdisciplinary Mathematics. Presently he is the President-in-Chief, Society of Statistics, Computer and Applications. He has also served as Editor, National Academy of Agricultural Sciences during July 2015 to December 2016.

6. Human resource development

VKG has been an outstanding teacher who took keen interest in Human Resource Development. He has taught several courses, which include Statistical Methods; Statistical Inference; Sampling Techniques; Design of Experiments; Linear Models, *etc.*; conducted a large number of ad-hoc Training Programmes under the aegis of Centre of Advanced Faculty Training, Summer/Winter Schools, for All India Coordinated Research Project on Energy Requirement in Agricultural Sector.

He has produced 50 (37 Ph.D. and 13 M.Sc.) students as Chairman/Co-Chairman Advisory Committee. These 50 students include 10 Ph.D. and 3 M.Sc. Students as Chairman Advisory Committee; 27 Ph.D. and 10 M.Sc. students as Co-Chairman Advisory Committee. A total of 42 distinct (08 both M.Sc. and Ph.D.) students have been mentored by him. Dr. Praggya Das is the first Ph.D. student guided by him as Chairman, Advisory Committee. Other 09 students guided by him as Chairman Advisory Committee are Drs. A. Dhandapani, Archana Pandey, D.V.V. Ramana, L.M. Bhar, Krishan Lal, J.S. Dhekale, Abhishek Rathore, Subhra Sarker, Subrata Kumar Satpati. Out of these Archana Pandey, D.V.V. Ramana, L.M. Bhar were also guided by him as Chairman Advisory Committee during their M.Sc. degree. He was also Co-Chairman Advisory Committee of Subrata Kumar Satpati. The other 32 students include

Sl. #	Name of Student	Degree
Co-Chairman (Chairman advisory Committee senior / colleague)		
1	Amitava Dey	Ph.D.
2	Ashish Das	Ph.D.
3	Cini Varghese	Ph.D.
4	Dileep Panda	Ph.D.
5	K. Venumadhav	Ph.D.
6	K. V. Palanichamy	Ph.D.
7	N. S. Gandhi Prasad	Ph.D.
8	Niti Agarwal	Ph.D.
9	P. K. Batra	Ph.D.
10	P. P. Yadav	Ph.D.
11	Priya Kohli	Ph.D.
12	Rajender Parsad	Ph.D.
13	Rajeshwar Singh	Ph.D.
14	Rama Krishna Singh	Ph.D.
15	Sanjay Dwivedi	M.Sc.
16	Seema Jaggi	M.Sc.
17	Somy Kuriakose	M.Sc.
18	Vijaya Raghava Kumar	Ph.D.
Co-Chairman (where Chairman, Advisory Committee is His students)		
19	Ajeet Kumar	Ph.D.
20	Ananta Sarkar	Ph.D.
21	B. N. Mandal	M.Sc., Ph.D.
22	Bishal Gurung	Ph.D.
23	Dinesh Pateria	Ph.D.
24	J. S. Tomar	Ph.D.
25	Jawaid Ashraf	M.Sc.
26	Pradip Nandi	M.Sc.
27	Ranjit Kumar Paul	M.Sc.
28	Sanklap Ojha	M.Sc.
29	Sarika	M.Sc., Ph.D.
30	Shashi Sekhar	M.Sc., Ph.D
31	Sukanta Dash	Ph.D.
32	Susheel Sarkar	M.Sc., Ph.D

VKG's students are working in academic, research and private organizations across the globe such as Former Director, Head of Divisions, National Fellow; Assistant Director General, ICAR; Scientists at ICAR Institutes; Faculty at State Agricultural Universities; Senior Researchers at International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) / International Maize and Wheat Improvement Center (CIMMYT); Professor at Indian Institute of Technology (IIT); Vice-Chancellor Management Institutes; Deputy Director General, Ministry of Statistics and Programme Implementation, Govt. of India; Advisor-in-Charge Monetary Policy Department, Reserve Bank of India; Professors in US Universities including SUNY Distinguished Professor; Senior Statisticians in Pharmaceuticals; HP, GENPact, Tata Consultancy Services, Banks, *etc.*

7. Significant contributions in design of experiments and sample surveys

During his long stint at ICAR-IASRI, VKG has made important and useful research contributions to both theoretical and applied aspects of Agricultural Statistics in general and Design of Experiments and Theory of Survey Sampling in particular. He has made outstanding theoretically creative and methodologically innovative contributions in the development of the subject of Statistics and its applications in the NARES. He always paid attention to the most challenging problems and came up with ingenious ideas that enhanced our understanding significantly. It is nearly impossible to describe all the research contributions of VKG in a limited space, and so, in what follows only a few examples of his major contributions are provided.

In an International Conference on Statistics and Informatics, organized by VKG as Organizing Secretary in 2006, Dr. M. S. Swaminathan in his written message said "*It is the effective use of the tools of statistical design of experiments that paved the way for the green revolution.*"

VKG has handled 25 research projects including ICAR National Fellow Scheme on '**Studies on Optimality of Designs Useful in Agricultural and Animal Experiments**' (1995-2000); ICAR National Professor Scheme '**Designs for Single Factor and Multi-factor Experiments and their Applications in Agricultural Systems Research**' (05.04.2006-16.03.2016) and externally funded projects from agencies like ICAR AP-Cess Fund, Department of Science and Technology (DST), Food and Agriculture Organization of the United Nations. In a FAO sponsored project, he has also contributed towards development of methodology for estimation of crop area and crop yield under mixed and continuous cropping.

7.1. Design of experiments

Some of his many notable contributions in design of experiments include obtaining optimal/efficient designs for various experimental settings *viz.* designs for making all possible pairwise treatment comparisons, designs for making test treatments-control comparisons, designs for diallel and tetra-allel crosses, block designs with nested factors, structurally complete/incomplete row-column designs, designs for bioassays, on-farm research, crop sequence experiments, estimation of competition effects, spatially correlated error structures, orthogonal fractional factorial plans, supersaturated designs and designs for microarray experiments. He has also significantly contributed towards investigating robustness of designs against several disturbances like loss of data, presence of outliers, model inadequacy, presence of trend, *etc.* He has also made immense contributions towards computer aided search of efficient designs for many experimental situations and for many practical ranges of parameters, shape and size of plots and blocks, diagnostics in designed field experiments; intercropping experiments; experiments with split application of inputs, *etc.*

Following new terminologies in Design of Experiments were introduced in the literature from research works of VKG and his collaborators: Generalized binary balanced block designs of type alpha; Balanced treatment incomplete block designs with unequal block sizes of Type G (for making test treatments control treatment comparisons under a heteroscedastic model when the intra block variances are proportional to integer power of block sizes); Balanced bipartite block designs with unequal block sizes; General efficiency balanced row-

column designs; Balanced bipartite row-column designs; Doubly nested partially balanced incomplete block designs; Structure resistant factorial designs; Cook and AP Statistics for detection of outliers in block designs.

7.2. Sample surveys

VKG has also made use of the combinatorics in sample surveys and because of this he has made several original contributions to the theory of survey sampling, particularly variance estimation from large scale complex survey data, unequal probability sampling, controlled selection and balanced and distance balanced sampling plans. He has also worked in the area of calibration approach of estimation of parameters. In fact, his work is pure genius with a real-world application. Nested stratified sampling; Distance balanced sampling plans; Inclusion probability proportional to size (IPPS) sampling plans excluding adjacent units, *etc.* are the new terminologies introduced by him and his collaborators. A brief of few of the above contributions is given in the sequel.

Variance Estimation based on Re-Sampling: Balanced Repeated Replication (BRR), a simple procedure for estimating variance of the estimator of a population parameter, introduced by McCarthy (1966) for a stratified design with number of units selected from each stratum as 2. It makes use of Hadamard Matrices for constructing "half-samples". This was extended to case of p units per stratum, whenever p is a prime number by Gurney and Jewett (1975). Gupta and Nigam (1987) established link between BRR and Mixed Orthogonal Arrays (MOA) of strength two and that allowed unequal number of selections from each stratum. They have also shown that MOA of Strength two, or equivalently, equal frequency orthogonal main effect plans for asymmetrical factorials give a set of balanced subsamples useful in variance estimation. Further, Dhandapani *et al.* (1996) have shown that certain proportional frequency plans can be used with no loss in efficiency for estimating the variance of a linear statistic. Some other references in this area include Magnussen (2009), Rao and Shao (1996), Rao and Shao (1999), Saigo *et al.* (2001) and many more.

Chakraborti (1963) introduced the use of incidence matrices of balanced incomplete block designs in sample selection. To understand an association between a combinatorial arrangement (design) and sampling plan, consider number of treatments (v) as population units (N); blocks (b) as samples; block size (k) as sample size (n) and treatments in blocks as units in the sample. Totality of all the blocks is sample space (S) and probability measure $p(s) = 1/b \; \forall s \in S$. The selection of a block is equivalent to selection of sample and the treatments in selected block denote the units in selected sample.

Chakraborti (1963) and Avadani and Sukhatme (1973) defined an association between Balanced Incomplete Block (BIB) Design vis-à-vis simple random sampling without replacement with $v = N, k = n, b =$ total number of samples, $r =$ number of times one unit is selected, $\lambda =$ number of times two units are in the same sample. The advantage of this association is that number of blocks b is much smaller than $\binom{N}{n}$.

Gupta *et al.* (1982) proposed a family of sampling schemes with inclusion probability proportional to size (IPPS) using the combinatorial properties of BIB designs. Denote X_i as the value of the auxiliary character for unit $i(i = 1, 2, ..., N)$ of the population, and $P_i = X_i/X$ where $X = X_1 + X_2 + ... + X_N$. The proposed plan selects a block s with

probability $p(s) = v[b(r-\lambda)]^{-1}[r\sum_{t\in s} P_t - \lambda]; s = 1, 2, ..., b$ of a BIB design with parameters $v = N, b, r, k \geq n, \lambda$ with notations having their usual meaning. They proved that if $k = n$ then the selected block constitute the required sample of the desired IPPS scheme. For $k \geq n$, they suggested selecting a subsample of size n by equal probability sampling without replacement from the k units belonging to the selected block s. They proved that the resulting plan has all the desirable properties of an IPPS sample.

Recently VKG worked in collaboration with Mandal and Rajender on balanced sampling plan excluding contiguous units. To understand the context, consider the example given by Hedayat *et al.* (1988): A researcher wants to estimate overall toxicity level of the soil at a toxic dump waste site. Over an extended period of time, the site is filled with 9 mounds of waste. Researcher has enough resources to sample 3 mounds. It is believed that due to seepage, the toxicity level of adjacent mounds will be similar. Obviously, researcher would try to avoid the selection of mounds that could have similar responses. When contiguous units in a population provide similar observations, then it is intuitively logical to avoid contiguous units in a sample. Naturally, simple random sampling without replacement (SRSWOR) is not appropriate in this case. A possible solution would be to consider a sampling plan that avoids the selection of adjacent units.

Under balanced sampling plan excluding contiguous units (BSA plans)

(i) π_i's are same for all units and

(ii) π_{ij}'s are same for all non-adjacent pairs of units and 0 for all adjacent pairs of units.

Here π_i and π_{ij} are the first order and second order inclusion probabilities. These plans are useful for sampling from populations where adjacent units are correlated because of some natural ordering in time or space.

Stufken *et al.* (1999) defined support of a BSA is related to Polygonal Designs (PD). A PD is an incomplete block design (ibd) such that

(i) each treatment has same number of replications and

(ii) every pair of non-adjacent treatments appear in same number of blocks and every pair of adjacent treatments never appear together in a block.

Mandal *et al.* (2008, 2011) presented construction methods and algorithms to obtain BSA plans / PDs. Mandal *et al.* (2008) extended BSA plans to IPPS plans excluding adjacent units (IPPSEA plans) where $\pi_i \propto p_i$, while adjacent pairs of units do not appear in any sample and presented characterizations of these plans and methods of obtaining these plans.

One drawback of BSA plans is that unbiased estimation of Narain-Horvitz-Thompson estimator of population mean is not possible since some second order inclusion probabilities are zero. To overcome this and improve BSA plans, Mandal *et al.* (2009) introduced distance balanced sampling plans (DBSPs). Under DBSPs, (i) π_i's are same for all units and (ii) two units with same distance have same value of π_{ij} and if the distance between units $(i, j) \geq$

distance between units (i', j') then $\pi_{ij} \geq \pi_{i'j'} \quad \forall i, i' \neq j, j' = 1, ..., N$. Mandal *et al.* (2009, 2011) presented construction methods and linear programming approach to obtain various classes of DBSPs. Gopinath *et al.* (2018) gave two-dimensional balanced sampling plans excluding adjacent units under sharing a border and island adjacency schemes.

It may be noted that no unit-by-unit selection method is available for controlled SRSWOR / IPPS plans, BSA, DBSP. Combinatorics involving experimental designs have interesting applications in survey sampling. Optimization techniques can be useful to obtain various classes of controlled sampling plans. All this work has appeared in a textbook entitled Combinatorics in Sample Surveys vis-à-vis Controlled Selection by LAP Lambert Academic Publishing GmbH & Co. KG, Verlag, Germany.

Nested stratified sampling: The procedure of dividing the population into strata and then again dividing each stratum into smaller strata and drawing a random sample from each one of the smaller strata may be called as nested stratified sampling. Nested stratified sampling enjoys all the advantages of stratified sampling *viz.* better management, administrative simplicity, representation of different sub-division of population as well as obtaining estimate of population parameter for each of the sub-divisions. It helps to deal with the situations in which another factor within each stratum influences the characteristics under study. For example, in crop yield estimation surveys for crop insurance, districts of community developments blocks CDB) may be strata and groups of homogeneous farmers within CDB are sub-strata.

Other research on survey sampling includes ratio estimators in two-phase sampling under size stratification and estimation over two successive occasions (with Pranesh Kumar in 1981) and calibration-based approach for product/regression type estimators in single/two phase with Sud and others (2014).

8. Prolific writer and collaborations

VKG has published 138 research papers in 40 International and National peer reviewed Journals (131 in Statistics Journals; 07 in Journals of Other Agricultural Sciences) that includes Biometrika, Technometrics, Journal of Royal Statistical Society, Series B; Applied Statistics (JRSS C); Australian Journal of Statistics; Journal of Statistical Planning and Inference; Statistics and Probability Letters; Sankhya; Metrika; Journal of Applied Statistics; Ars Combinatoria; Utilitas Mathematica; Communications in Statistics: Theory and Methods; Biometrical Journal; Journal of Statistical Theory and Practice; Journal of the Indian Society of Agricultural Statistics, Better Crops, Indian Journal of Agronomy, Agricultural Research, *etc.*

He has written 05 text books; produced 02 electronic books; 05 Research monographs; 04 Technical Bulletins; 12 Book Chapters; 21 Project/Technical Reports; 03 Workshop Proceedings; 11 Volumes of various Journals as Guest Editor; 02 Book Reviews. He has 08 copyrights to his credit.

VKG has been a very active partner in developing statistical packages for generation of randomized layout of designs and also their analysis *viz.* Statistical Package for Balanced Incomplete Block Designs; Statistical Package for Factorial Experiments; Statistical Package for Augmented Designs; Statistical Package for Repeated Measurement Designs, *etc.* He has

also contributed significantly towards web portals "Sample Survey Resources Server" as Team Leader and "Design Resources Server" as Collaborator, which are being viewed throughout the globe.

His commitment and energy have benefitted his wide range of collaborations. He has 106 research collaborators spread across 22 organizations: 04 from US, one from Japan; 02 from CGIAR Organizations; ISI Kolkata, 05 Traditional Universities in India, 02 SAUs; 06 ICAR Institutes including IASRI. His research has received wide acceptance and appreciation throughout the globe. His research papers have been cited widely including 08 text books published by Elsevier Science, Springer Verlag, John Wiley & Sons, North Holland, Hindustan Book Agency.

The complete list of publications is given at the end before references. The references that are part of list of publications have not been given to avoid repetition.

9. Good listener and amenable to suggestions

I would like to mention here that along with his research, mentoring and managerial capabilities, VKG is good listener even to his students and juniors and filters alone the good points and adopts them

साधु ऐसा चाहिए, जैसा सूप सुभाय।
सार-सार को गहि रहै, थोथा देई उड़ाय॥

I would like to mention one success story. In 1998, he was Course Director for a Centre of Advanced Studies Training [now known as Centre of Advanced Faculty Training] programme. During the discussions, we the junior colleagues requested him that for spreading awareness about the advances in Statistical Techniques among agricultural researchers for adoption we should invite Agricultural scientists from other disciplines along with statisticians. He listened to us and implemented. Perhaps that was the first ad-hoc Training Programme (I stand corrected if I am wrong) in which Agricultural scientists from other disciplines along with statisticians were invited and the training was successful. Since then, in most of the training programmes, this symbiosis is being maintained resulting into adoption of advanced statistical techniques in NARES.

I have made an attempt to just present a brief. I have no words at my command to express the vast contributions made by him.

सब धरती कागज करूँ, लिखनी सब बनराय।
सात समुद्र की मसि करूँ, गुरु गुण लिखा न जाय॥

Besides that, outreach training programmes, interaction with agricultural researchers led to adoption of efficient designs and advanced statistical techniques. Concerted efforts made by VKG and his collaborators in disseminating advance statistical techniques to the stakeholders in NARES has led to the adoption of efficient designs like balanced incomplete block (BIB) designs, extended group divisible (EGD) designs, designs for fitting response surfaces, resolvable block designs for experimentation and sophisticated analytics of the data generated by the researchers. The linear programming approach has been used for obtaining energy requirements in agricultural sector.

10. List of published papers: 138

1. Mahesh Kumar Panda, Tofan Kumar Biswal, and V. K. Gupta (2025). Bayesian A-optimal designs for Gamma and Poisson regression models: An algorithmic approach. *Journal of the Indian Society of Agricultural Statistics*, **79**, 63-78.
2. Mahesh Kumar Panda, Tofan Kumar Biswal, and V. K. Gupta (2025). R-optimal designs for Gamma regression model with two parameters. *Statistics and Applications*, **23**, 33-53.
3. V. K. Gupta, Bikas. K. Sinha, and Bimal K. Sinha (2024). List of 103 selected research papers of C. R. Rao. *Statistics and Applications*, **22** (Special Issue in the Fond Memories of C R Rao on "Life and Work of CR Rao (1920-2023): The Revolutionary of Statistical Sciences"), 37-43.
4. B. N. Mandal and V. K. Gupta (2023). Linear integer programming and its innovative applications in design of experiments and sample surveys. *Journal of the Indian Society of Agricultural Statistics*, **77**, 35-42
5. V. K. Gupta, Rahul Mukerjee, Mausumi Bose, and Arun K. Nigam (2021). Tribute to Professor Aloke Dey – A noble and pious soul. *Statistics and Applications*, **19**, 1-10.
6. V. K. Gupta and Rahul Mukerjee (2019). Professor Arun Kumar Nigam: A tribute. *Statistics and Applications*, **17**, 1-10
7. V. K. Gupta and Rahul Mukerjee (2018). J.N.K. Rao – A tribute. *Statistics and Applications*, **16**, 1-19.
8. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2017). Balanced treatment incomplete block designs through integer programming. *Communications in Statistics: Theory and Methods*, **46**, 3728-3737.
9. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2016). Cyclic circular balanced and strongly balanced cross over designs through integer programming. *Communications in Statistics: Theory and Methods*, **45**, 859-871.
10. P. K. Nandi, Rajender Parsad, and V. K. Gupta (2015). Simultaneous optimization of incomplete multi-response experiments. *Open Journal of Statistics*, **5**, 430-44. http://dx.doi.org/10.4236/ojs.2015.55045
11. S. Dash, Rajender Parsad, and V. K. Gupta (2014). Efficient row-column designs with two rows. *Journal of the Indian Society of Agricultural Statistics*, **68**, 377-390.
12. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2014). Efficient incomplete block designs through linear integer programming. *American Journal of Mathematical and Management Sciences*, **33**, 110-124.
13. Shashi Shekhar, Lalmohan Bhar, and V. K. Gupta (2014). Incomplete block designs for multiple asymmetric parallel line assays. *International Journal of Computational and Theoretical Statistics*, **1**, 29-35.
14. U. C. Sud, Hukum Chandra, and V. K. Gupta (2014). Calibration based product estimator in single- and two-phase sampling. *Journal of Statistical Theory and Practice*, **8**, 1-11.

15. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2014). Construction of efficient multi-level k-circulant supersaturated designs. *Communications in Statistics - Theory and Methods*, **43**, 599-615.

16. U. C. Sud, Hukum Chandra, and V. K. Gupta (2013). Calibration approach based regression type estimator for inverse relationship between study and auxiliary variable. *Journal of Statistical Theory and Practice*, **8**, 707-721.

17. Rohan Kumar Raman, U. C. Sud, Hukum Chandra, and V. K. Gupta (2013). Calibration estimator of population total with sub-sampling of non-respondents. *Journal of the Indian Society of Agricultural Statistics*, **67**, 329-337.

18. Sukanta Dash, Rajender Parsad, and V. K. Gupta (2013). Row-column designs for 2n factorial 2-colour microarray experiments for estimations of main effects and two factor interactions for orthogonal parameterization. *Agricultural Research*, **2**, 172-182.

19. Lalmohan Bhar, V. K. Gupta, and Rajender Parsad (2013). Detection of outliers in designed experiments in presence of masking. *Statistics and Applications*, **11**, 147-160.

20. V. K. Singh, V. Govil, S. K. Singh, B. S. Dwivedi, M. C. Meena, V. K. Gupta, K. Majumdar, and B. Gangwar (2012). Precision nutrient management strategies using GIS-based mapping in Western Uttar Pradesh. *Better Crops, South Asia*, **6**, 15-18.

21. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2012). Generation of binary incomplete block design with a specified concurrence matrix. *Journal of Statistics and Applications : A Journal of 'Forum for Interdisciplinary Mathematics'*, **7**, 9-25.

22. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2012). Doubly nested partially balanced incomplete block designs. *Journal of Statistics and Applications : A Journal of 'Forum for Interdisciplinary Mathematics'*, **7**, 41-57.

23. U. K. Pradhan, Krishan Lal, and V. K. Gupta (2012). Optimum conditions for mixture experiments with process variable for the expected response with minimum variability. *Statistics and Applications*, **10**, 63-71.

24. V. K. Gupta and Rajender Parsad (2012). Use of incomplete block designs in agricultural experiments. *Indian Journal of Agronomy*, **57**, 107-116.

25. Subrata Kumar Satpati, V. K. Gupta, Rajender Parsad, and M. L. Aggarwal (2012). Computer generated change over designs for correlated observations. *Communications in Statistics: Theory and Methods*, **41**, 3786-3798.

26. V. K. Gupta, Kashinath Chatterjee, Ashish Das, and Basudev Kole (2012). Addition of runs to an s-level supersaturated design. *Journal of Statistical Planning and Inference*, **142**, 2402-2408.

27. Krishan Lal, Rajender Parsad, and V. K. Gupta (2012). Robustness of block designs on the basis of pairwise treatment comparisons against missing data. *Model Assisted Statistics and Applications*, **7**, 75-80: DOI: 10.3233/MAS-2011-0213.

28. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2011). Construction of efficient mixed-level k-circulant supersaturated designs. *Journal of Statistical Theory and Practice*, **5**, 627-648.

29. V. K. Gupta, A. K. Nigam, Rajender Parsad, L. M. Bhar, and Subrat Keshori Behera (2011). Resolvable block designs for factorial experiments with full main effects efficiency. *Journal of the Indian Society of Agricultural Statistics*, **65**, 305-315.

30. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2010). Construction of polygonal designs using linear integer programming. *Communications in Statistics – Theory and Methods*, **40**, 1787-1794.

31. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2010). Inclusion probability inversely proportional to size sampling. *Journal of Statistics and Applications : A Journal of 'Forum for Interdisciplinary Mathematics'*, **5**, 15-26.

32. V. K. Gupta, Poonam Singh, Basudev Kole, and Rajender Parsad, (2010). Addition of runs to a two-level supersaturated design. *Journal of Statistical Planning and Inference*, **140**, 2531-2535.

33. V. K. Gupta, Poonam Singh, Basudev Kole, and Rajender Parsad (2010). Construction of efficient balanced and nearly balanced two-level supersaturated designs. *Journal of Statistics and Applications : A Journal of 'Forum for Interdisciplinary Mathematics'*, **5**, 179-194.

34. Basudev Kole, Jyoti Gangwani, V. K. Gupta, and Rajender Parsad (2010). Two-level supersaturated designs: A review. *Journal of Statistical Theory and Practice*, **4**, 589-608.

35. V. K. Gupta, Poonam Singh, Basudev Kole, and Rajender Parsad (2010). Computer aided construction of efficient multi-level supersaturated designs. *Journal of Statistical Theory and Practice*, **4**, 221-231.

36. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2010). Linear integer programming approach to construct distance balanced sampling plans. *Journal of the Indian Society of Agricultural Statistics*, **64**, 303-312.

37. Ananta, Rajender Parsad, V. K. Gupta, Kashinath Chatterjee, and Abhishek Rathore (2010). Efficient row-column designs for micro-array experiments. *Journal of the Indian Society of Agricultural Statistics*, **64**, 89-117.

38. B. N. Mandal, Rajender Parsad and V. K. Gupta, and U. C. Sud (2009). A family of distance balanced sampling plans. *Journal of Statistical Planning and Inference*, **139**, 860-874.

39. V. K. Gupta, Poonam Singh, Basudev Kole, and Rajender Parsad (2009). Construction of efficient unbalanced mixed-level supersaturated designs. *Statistics and Probability Letters*, **79**, 2369-2366.

40. V. K. Gupta, L. M. Bhar, Basudev Kole, and Rajender Parsad (2009). Efficient unbalanced mixed-level supersaturated designs. *Kageyama Special Issue of Journal of Statistics and Applications, Journal of Forum for Interdisciplinary Mathematics*, **4**, 443-445.

41. Rajender Parsad, Abhishek Rathore, and V. K. Gupta (2009). Computer aided construction of efficient designs for making treatment-treatment and treatment-control comparisons. *Special Issue of the American Journal of Mathematical and Management Sciences in Memory of Bechhofer-Gupta-Sobel*, **29**, 201-228.

42. V. K. Gupta, Poonam Singh, Basudev Kole, and Rajender Parsad (2009). Construction of optimal mixed-level supersaturated designs. *Journal of the Indian Society of Agricultural Statistics*, **63**, 311-319.

43. Susheel Kumar Sarkar, Krishan Lal, Rajender Parsad, and V. K. Gupta (2009). Generation of linear trend free designs for factorial experiments. *Journal of the Indian Society of Agricultural Statistics*, **63**, 165-174.

44. Cini Varghese, A. R. Rao, and V. K. Gupta (2009). Optimal change over designs under treatment × unit interaction. *International Journal of Applied Mathematics and Statistics*, **14**, 27-34.

45. Rajender Parsad, V. K. Gupta, Jitendra Kumar, and N. A. Shakil (2008). Significance of statistical tools for agrochemical research. *Pesticide Research Journal*, **20**, 8-17.

46. V. K. Gupta, Rajender Parsad, L. M. Bhar, and Basudev Kole (2008). Supersaturated designs for asymmetrical factorial experiments. *Journal of Statistical Theory and Practice*, **2**, 95-108.

47. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2008). IPPS sampling plans excluding adjacent units. *Communications-in-Statistics: Theory and Methods*, **37**, 2532-2550.

48. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2008). Computer aided construction of balanced sampling plans excluding contiguous units. *Journal of Statistics and Applications: A Journal of 'Forum for Interdisciplinary Mathematics'*, **3**, 59-85.

49. V. K. Gupta (2008). Orthogonal arrays and their applications. *Journal of the Indian Society of Agricultural Statistics*, **62**, 1-18.

50. B. N. Mandal, Rajender Parsad, V. K. Gupta, and U. C. Sud (2008). Nested stratified sampling. *Model Assisted Statistics and Applications*, **3**, 211-223.

51. V. K. Gupta, Rajender Parsad, L. M. Bhar, and Basudev Kole (2008). Computer-generated construction of efficient two level supersaturated designs. *Journal of the Indian Society of Agricultural Statistics*, **62**, 183-194.

52. R. Srivastava, Rajender Parsad, Amitava Dey, and V. K. Gupta. (2008). A-efficient block designs for multiple parallel line assays. *Journal of the Indian Society of Agricultural Statistics*, **62**, 231-243.

53. Rajender Parsad, P. K. Nandi, V. K. Gupta, and L. M. Bhar (2008). Outliers in multi-response experiments. *Statistics and Applications*, **6**, 275-292.

54. Subrata Kumar Satpati, Rajender Parsad, and V. K. Gupta (2007). Efficient block designs for dependent observations: A computer aided search. *Communications-in-Statistics: Theory & Methods*, **36**, 1187-1223.

55. Rajender Parsad, V. K. Gupta, and R. Srivastava (2007). Designs for cropping systems research. *Journal of Statistical Planning and Inference*, **137**, 1687-1703.

56. R. Srivastava, Rajender Parsad, Amitava Dey, and V. K. Gupta (2007). A-efficient block designs for symmetric parallel line assays. *Utilitas Mathematica*, **73**, 239-253.

57. Rajender Parsad, Sanpei Kageyama, and V. K. Gupta (2007). Use of complementary property of block designs in PBIB designs. *Ars Combinatoria*, **85**, 173-182.

58. Seema Jaggi, Cini Varghese, and V. K. Gupta (2007). Optimal circular block designs for neighbourig competition effects. *Journal of Applied Statistics*, **34**, 577-584.

59. Anu Sharma, Cini Varghese, A. R. Rao, V. K. Gupta, and Soumen Pal (2007). SPRMD – A statistical package for cataloguing and generation of repeated measurements designs. *Journal of the Indian Society of Agricultural Statistics*, **61**, 84-90.

60. Ananta Sarkar, Rajender Parsad, Abhishek Rathore, and V. K. Gupta (2007). Efficient block designs for microarray experiments. *Journal of the Indian Society of Agricultural Statistics*, **61**, 227-248.

61. Krishan Lal, Rajender Parsad, and V. K. Gupta (2007). Trend-free nested balanced incomplete block designs and designs for diallel cross experiments. *Calcutta Statistical Association Bulletin*, **59**, 203-221.

62. Subrata Kumar Satpati, Rajender Parsad, V. K. Gupta, and A.K. Nigam (2006). Computer aided search of efficient nested incomplete block designs for correlated observations. *Journal of Combinatorics, Information and System Sciences (Special issue in Honour of Prof. Bhu Dev Sharma)*, **31**, 163-186.

63. Abhishek Rathore, Rajender Parsad, and V. K. Gupta (2006). Computer aided search of efficient block designs for making all possible pairwise treatment comparisons. *Journal of Statistics and Applications: A Journal of 'Forum for Interdisciplinary Mathematics'*, **1**, 15-33.

64. R. Srivastava, Rajender Parsad, Amitava Dey, and V. K. Gupta (2006). A-efficient block designs for asymmetrical slope ratio assays. *Journal of Statistics and Applications: A Journal of 'Forum for Interdisciplinary Mathematics'*, **1**, 185-192.

65. Seema Jaggi, V. K. Gupta, and Jawaid Ashraf (2006). On block designs partially balanced for neighbouring competition effects. *Journal of the Indian Statistical Association*, **44**, 27-41.

66. Subhra Sarker, Rajender Parsad, and V. K. Gupta (2005). Outliers in block designs for diallel crosses. *Metron-International Journal of Statistics*, **63**, 177-191.

67. Rajender Parsad, V. K. Gupta, and Sudhir Gupta (2005). Optimal designs for experiments on two-line and four-lines crosses. *Utilitas Mathematica*, **68**, 11-32.

68. Abhishek Rathore, Rajender Parsad, and V. K. Gupta (2004). Computer aided construction and analysis of augmented designs. *Journal of the Indian Society of Agricultural Statistics*, **57** (special volume), 320-344.

69. Sangeeta Ahuja, Rajender Parsad, V. K. Gupta, and P. K. Malhotra (2004). Statistical package for factorial experiments. *Journal of the Indian Society of Agricultural Statistics*, **58**, 312-330.

70. Rajender Parsad, V. K. Gupta, and N.S.G. Prasad (2003). Structurally incomplete row-column designs. *Communication in Statistics: Theory and Methods*, **32**, 239-261.

71. S. Sarker, V. K. Gupta, and Rajender Parsad (2003). Robust block designs for making test treatments – control treatment comparisons against the presence of an outlier. *Journal of the Indian Society of Agricultural Statistics*, **56**, 7-18.

72. Krishan Lal, V. K. Gupta, and L. M. Bhar (2003). Robustness of Youden square designs against missing data. *Journal of the Indian Society of Agricultural Statistics*, **56**, 107-116.

73. J. S. Dhekale, Rajender Parsad, and V. K. Gupta (2003). Analysis of intercropping experiments by using experiments with mixtures methodology. *Journal of the Indian Society of Agricultural Statistics*, **56**, 260-267.

74. V. K. Gupta, Rajender Parsad, and Ashish Das (2003). Variance balanced block designs with unequal block sizes. *Utilitas Mathematica*, **64**, 183-192.

75. V. K. Gupta, D. V. V. Ramana, and Rajender Parsad (2002). Weighted A-optimal block designs for comparing test treatments with controls with unequal precision. *Journal of Statistical Planning and Inference (Special issue in the memory of Professor Yamamoto)*, **106**, 159-175.

76. L. M. Bhar and V. K. Gupta (2002). Robust row-column designs for complete diallel cross experiments. *Metrika*, **56**, 83-90.

77. Krishan Lal, V. K. Gupta, and L. M. Bhar (2001). Robustness of designed experiments against missing data. *Journal of Applied Statistics*, **28**, 63-79.

78. Rajender Parsad, V. K. Gupta, and Daniel Voss (2001). Optimal nested row-column designs. *Journal of the Indian Society of Agricultural Statistics*, **54**, 244-257.

79. V. K. Gupta and Rajender Parsad (2001). Block designs for comparing test treatments with control treatments - An overview. *Special Issue of Statistics and Applications to felicitate the 80th Birthday of Dr. M.N. Das*, **3**, 133-146.

80. Rajender Parsad and V. K. Gupta (2001). Balanced bipartite row-column designs. *Ars Combinatoria*, **61**, 301-312.

81. Seema Jaggi, V. K. Gupta, and V. K. Sharma (2001). Design and analysis of agroforestry experiments: An overview. *Journal of the Indian Society of Agroforestry*, **3**, 120-129.

82. L. M. Bhar and V. K. Gupta (2001). A useful statistic for studying outliers in experimental designs. *Sankhya*, **63**, 338-350.

83. R. Srivastava, Rajender Parsad, and V. K. Gupta (2000). Structure resistant factorial designs. *Sankhya*, **B62**, 257-265.

84. Rajender Parsad and V. K. Gupta (2000). A note on augmented designs. *Indian Journal of Plant Genetic Resources*, **13**, 53-58.

85. V. K. Gupta, D. V. V. Ramana, and Rajender Parsad (1999). Weighted A-efficiency of block designs for making treatment-control and treatment-treatment comparisons. *Journal of Statistical Planning and Inference*, **77**, 4301-4320.

86. Rajender Parsad, V. K. Gupta, and R. Srivastava (1999). Universally optimal block Designs For Diallel Crosses. *Statistics and Applications*, **1**, 35-52.

87. Seema Jaggi, Rajender Parsad, and V. K. Gupta (1999). Construction of non-proper balanced bipartite block designs. *Calcutta Statistical Association Bulletin*, **49**, 55-63.

88. P. K. Batra, Rajender Parsad, V. K. Gupta, and O. P. Khanduri (1999). An alternate strategy for design and analysis of experiments involving split application of fertilizer. *Statistics and Applications*, **1**, 175-187.

89. V. K. Gupta, Archana Pandey, and Rajender Parsad (1998). A-optimal block designs under mixed model for making test treatments control comparisons. *Sankhya*, **B 60**, 496-510.

90. V. K. Gupta, D. V. V. Ramana, and S. K. Agarwal (1998). Weighted A-optimal row-column designs for making treatment-control and treatment-treatment comparisons. *Journal of Combinatorics, Information and System Sciences*, **23**, 333-344 (Special Issue in honour of Professor J.N. Srivastava).

91. Seema Jaggi, Rajender Parsad, and V. K. Gupta (1997). General efficiency balanced designs with unequal block sizes for comparing two sets of treatments. *Journal of the Indian Society of Agricultural Statistics*, **50**, 37-46.

92. Seema Jaggi and V. K. Gupta (1997). A-optimal block designs with unequal block sizes for comparing two disjoint sets of treatments. *Sankhya*, **B 59**, 164-180.

93. Seema Jaggi, V. K. Gupta, and Rajender Parsad (1996). A-efficient block designs for comparing two disjoint sets of treatments. *Communication in Statistics: Theory and Methods*, **25**, 967-983.

94. R. Srivastava, Rajender Parsad, and V. K. Gupta (1996). Robustness of block designs for making test treatments - control comparisons against a missing observation. *Sankhya*, **B 58**, 407-413.

95. Rajender Parsad, V. K. Gupta, and V. P. N. Singh (1996). Trace optimal designs with unequal block sizes for comparing two disjoint sets of treatments. *Sankhya*, **B 58**, 414-426.

96. A. Dhandapani, V. K. Gupta, and A. K. Nigam (1996). Variance estimation using proportional frequency plans. *Golden Jubilee Volume of Journal of the Indian Society of Agricultural Statistics*, **49**, 267-276.

97. Rajender Parsad, V. K. Gupta, and N. S. G. Prasad (1995). On construction of A-efficient balanced test treatments incomplete block designs. *Utilitas Mathematica*, **47**, 185-190.

98. V. K. Gupta (1995). Universally optimal generalized binary balanced block designs of type alpha. *Sankhya*, **B 57**, 420-427.

99. Rajender Parsad and V. K. Gupta (1994). A optimality of group divisible designs with unequal block sizes for making test treatment control comparisons under a heteroscedastic model. *International Journal of Mathematical and Statistical Sciences*, **3**, 5-31.

100. Rajender Parsad and V. K. Gupta (1994). Optimal block designs with unequal block sizes for treatment control comparisons under a heteroscedastic model. *Sankhya*, **B 56**, 449-461.

101. G. C. Chawla and V. K. Gupta (1993). Two-factor change over designs based on William's squares. *International Journal of Applied Statistics*, 326-334.

102. V. K. Gupta and D. V. V. Ramana (1993). Robustness of designs against a single missing observation. *International Journal of Mathematical and Statistical Sciences*, **2**, 45-56.

103. V. K. Gupta (1993). Optimal nested block designs. *Journal of the Indian Society of Agricultural Statistics*, **45**, 187-194.

104. V. K. Gupta and R. Srivastava (1992). Investigations on robustness of block designs against missing observations. *Sankhya*, **B 54**, 100-105.

105. Ashish Das, V. K. Gupta, and P. Das (1992). E-optimal block designs under heteroscedastic model. *Communication in Statistics: Theory & Methods*, **21**, 1651-1666.

106. R. Srivastava, V. K. Gupta, and A. Dey (1991). Robustness of some designs against missing data. *Journal of Applied Statistics*, **18**, 303-308.

107. V. K. Gupta, Ashish Das, and A. Dey (1991). Universal optimality of block designs with unequal block sizes. *Statistics and Probability Letters*, **11**, 177-180.

108. Rajeshwar Singh and V. K. Gupta (1991). Resistance of block designs. *Journal of Statistical Planning and Inference*, **27**, 263-269.

109. S. K. Agarwal, V. K. Gupta, and B. V. S. Sisodia (1991). A study of performance of an alternative estimator of population total for unequal probability sampling procedures. *Cahiers du CERO*, **33**, 71-78.

110. A. Dey, R. Srivastava, and V. K. Gupta (1991). Robust designs-A review and bibliography. *Cahiers du CERO*, **33**, 51-62.

111. Rajeshwar Singh and V. K. Gupta (1991). E-optimal designs for two-way elimination of heterogeneity. *Sankhya*, **B 53**, 97-104.

112. V. K. Gupta and N. S. G. Prasad (1991). On construction of general efficiency balanced block designs. *Sankhya*, **B 53**, 89-96.

113. R. Srivastava, V. K. Gupta, and A. Dey (1990). Robustness of some designs against missing observations. *Communication in Statistics: Theory & Methods*, **191**, 121-126.

114. V. K. Gupta and N. S. G. Parsad (1990). On general efficiency balanced row-column designs. *Calcutta Statistics Association Bulletin*, **39**, 172-183.

115. V. K. Gupta and Rajeshwar Singh (1990). Characterization and construction of E-optimal block designs. *Sankhya*, **B 52**, 204-211.

116. V. K. Gupta (1989). A method of inclusion probability proportional to size selection. *Statistics*, **49**, 277-281.

117. V. K. Gupta and Rajeshwar Singh (1989). On E-optimal block designs. *Biometrika*, **76**, 184-188.

118. V. K. Gupta and Rajeshwar Singh (1988). A note on balanced block designs *Journal of the Indian Society of Agricultural Statistics*, **40**, 45-50.

119. V. K. Gupta, A. K. Nigam, and P. D. Puri (1987). Characterization and construction of incomplete block designs for symmetrical parallel line assays. *Journal of the Indian Society of Agricultural Statistics*, **39**, 161-166.

120. V. K. Gupta and A. K. Nigam (1987). Mixed orthogonal arrays for variance estimation when number of primary selections is unequal in each stratum. *Biometrika*, **74**, 735-742.

121. Pranesh Kumar, V. K. Gupta, and A. K. Nigam (1987). On sample selection procedures using two auxiliary characters. *Statistics*, **47**, 219-225.

122. A. Dey and V. K. Gupta (1986). Another look at the efficiency and partially efficiency balanced designs. *Sankhya*, **B 48**, 437-438.

123. Pranesh Kumar, V. K. Gupta, and S. K. Agarwal (1985). On variance estimation in unequal probability sampling. *Australian Journal of Statistics*, **27**, 195-201.

124. A. K. Nigam and V. K. Gupta (1985). Construction of orthogonal main-effect plans using Hadamard matrices. *Technometrics*, **27**, 37-48.

125. Pranesh Kumar, V. K. Gupta, and A. K. Nigam (1985). On inclusion probability proportional to size sampling scheme. *Journal of Statistical Planning and Inference*, **12**, 127-131.

126. V. K. Gupta and A. K. Nigam (1985). A class of asymmetrical orthogonal resolution IV designs. *Journal of Statistical Planning and Inference*, **12**, 381-383.

127. A. K. Nigam, Pranesh Kumar, and V. K. Gupta (1984). Some methods of inclusion probability proportional to size sampling. *Journal of the Royal statistical Society*, **B 46**, 564-571.

128. A. K. Nigam and V. K. Gupta (1984). A method of sampling with equal or unequal probabilities without replacement. *Applied Statistics, (JRSS C)*, **33**, 227-229.

129. A. Dey, V. K. Gupta, and M. Singh (1983). Optimal change over designs. *Sankhya*, **B 45**, 223-239.

130. A. K. Nigam, V. K. Gupta, and A. C. Bora (1982). Four factor balanced optimal designs of resolution V in terms of a modified trace criterion. *Biometrical Journal*, **24**, 425-427.

131. V. K. Gupta, A. K. Nigam, and Pranesh Kumar (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, **69**, 191-196.

132. V. K. Gupta, A. K. Nigam, and A. Dey (1982). Orthogonal main effect plans for asymmetrical factorials. *Technometrics*, **24**, 135-137.

133. V. K. Gupta and A.K. Nigam (1982). On a model useful for approximating fertiliser-response relationships. *Journal of the Indian Society of Agricultural Statistics*, **34**, 61-74.

134. Pranesh Kumar and V. K. Gupta (1981). On ratio estimators in two-phase sampling under size stratification and estimation over two successive occasions. *Journal of Mathematical Operations Forsch Statistics, Series Statistics*, **12**, 409-417.

135. Pranesh Kumar and V. K. Gupta (1981). On ratio estimators in two-phase sampling under size stratification. *Journal of Mathematical Operations Forsch Statistics, Series Statistics*, **12**, 419-427.

136. M. Singh and V. K. Gupta (1980). On the estimation of mean parameter with random number of observations on each unit. *Biometrical Journal*, **22**, 41-45.

137. I. J. Singh and V. K. Gupta (1977). Predicting response of wheat to fertilizer use under uncertain fertility. *Haryana Agricultural University Journal Research*, **6**, 52-60.

138. V. K. Gupta and Pranesh Kumar (1977). Study of agricultural efficiency in Haryana. *Haryana Agricultural University Journal Research*, **7**, 119-125.

11. List of books published: 05

1. Raosaheb Latpate, Jayant Kshirsagar, V. K. Gupta, and Girish Chandra (2021). *Advanced Sampling Methods*. Springer Singapore. ISBN: 978-981-16-0621-2, https://doi.org/10.1007/978-981-16-0622-9, XVII, 301.
2. V. K. Gupta, Rajender Parsad, Lal Mohan Bhar, and B. N. Mandal (2016). *Statistical Analysis of Agricultural Experiments – Part-I: Single Factor Experiments*. ICAR-IASRI, Publication Sponsored by Education Division, ICAR.
3. V. K. Gupta, B. N. Mandal, and Rajender Parsad (2012). *Combinatorics in Sample Surveys vis-a-vis Controlled Selection*. LAP Lambert Academic Publishing GmbH and Co. KG, Verlag, Germany.
4. A. K. Nigam, P. D. Puri, and V. K. Gupta (1989). *Characterization and Analysis of Block Design*. Wiley Eastern Limited.
5. A. K. Nigam and V. K. Gupta (1979). *A Handbook of Analysis of Agricultural Experiments*. Published by IASRI, New Delhi.

12. List of monographs published: 05

1. V. K. Gupta, U. C. Sud, and Chandrabhanu Chauhan (2015). *Weighing and Calibration in Sample Survey Estimation*. ICAR-IASRI, New Delhi. Publication Sponsored by Education Division, ICAR.
2. V. K. Gupta, A. Dhandapani, and Rajender Parsad (2007). *Monograph on Hadamard Matrices* (along with a CD-ROM). ICAR-IASRI, New Delhi. Publication Sponsored by Education Division, ICAR. [64 pages along with a CD-ROM].
3. Rajender Parsad, V. K. Gupta, P. K. Batra, S. K. Satpati, and Pabitra Biswas (2007). *Monograph on α-Designs*. ICAR-IASRI, New Delhi. [174 pages].
4. A. K. Nigam and V. K. Gupta (1985). *Optimality of Multi-Dimensional Designs*. IASRI, New Delhi.
5. A. K. Nigam, V. K. Gupta, and Prem Narain (1981). *Supplemented Block Designs*. IASRI, New Delhi.

13. List of technical bulletins published: 04

1. V. K. Gupta, Rajender Parsad, and B. N. Mandal (2015). *Significance of Experimental Designs in Agricultural Research*. ICAR-IASRI, New Delhi. Publication sponsored by Education Division, ICAR IASRI/TB-01/2015.
2. Rajender Parsad and V. K. Gupta (2009). Design Resources Server. ICAR-IASRI, New Delhi. I.A.S.R./T.B.-03/2009.
3. V. K. Gupta, Krishan Lal, and Rajender Parsad (2008). *Experiments with Mixtures: A Bibliography*. IASRI, New Delhi. I.A.S.R.I./T.B.-02/2008.
4. V. K. Gupta, Rajender Parsad, and Seema Jaggi (2001). *Division of Design of Experiments: 1990-2000*. IASRI Publication I.A.S.R.I./T.B.-01/2001.

14. List of electronic books: 02

1. *Design and Analysis of Agricultural Experiments*. Compiled, Edited and Developed by Rajender Parsad, R. Srivastava, and V. K. Gupta. This book is divided into 5 modules and contains 63 lectures described in 736 pages.
2. *Advances in Data Analytical Techniques*. Compiled, Edited and Developed by Rajender Parsad, V. K. Gupta, L. M. Bhar, and V. K. Bhatia. This E-book consists of chapters on 66 topics described in more than 920 pages. The topics covered in this electronic book have been categorized into six modules namely Computer Usage and Statistical Software Packages, Basic Statistical Techniques consisting of Statistical methods and Inference, Design of Experiments and Sample Surveys, Diagnostics and Remedial Measures, Applications of Multivariate Techniques, Modelling and Forecasting Techniques in Agriculture and Other Useful Techniques such as Bio-informatics, Geoinformatics, Microarrays, DNA Fingerprinting, nanotechnology in agriculture *etc*.

15. List of book chapters published: 12

1. Krishan Lal, Rajender Parsad, and V. K. Gupta (2015). Trend free block designs under a heteroscedastic model. In: *Statistics and Informatics in Agricultural Research*, Indian Society of Agricultural Research, New Delhi (Eds. U. C. Sud, Hukum Chandra, L. M. Bhar, Susheel Kumar Sarkar)
2. B. N. Mandal, V. K. Gupta, and Rajender Parsad (2015). Distance Balanced Sampling Plans - An Overview. In *Statistical and Mathematical Sciences and their Applications*, Narosa Publishing House Pvt. Ltd. (Ed. Neeraj Tiwari, Kumayun University), 90-99.
3. Nitiprasad N. Jambhulkar, Krishan Lal, Rajender Parsad, and V. K. Gupta (2015). Minimum aberration fractional factorial plans for multi-level experiments. In *Statistical and Mathematical Sciences and their Applications*, Narosa Publishing House Pvt. Ltd. (Ed: Neeraj Tiwari, Kumayun University), 41-73.
4. Prem Narain, V. K. Gupta, and Rajender Parsad (2015). Significance of Statistical Sciences in Agricultural Research. In *100 Years of Agricultural Research*, NAAS Publication.
5. V. K. Gupta, Rajender Parsad, and Krishan Lal (2011). Design and Analysis of Animal Nutrition Experiments. In *Animal Nutrition: Advancements in Feeds and Feeding of Livestock*, Agrobios (India), (Eds: Lokesh Gupta and K.K. Singhal, Jodhpur), 425-448.
6. Rajender Parsad, Jose Crossa, V. K. Gupta, Raj K. Gupta, J. K. Ladha, and Raman K. Anitha (2009). Statistical tools for farmers' participatory trials for conservation agriculture. In *Integrated Crop and Resource Management in the Rice-Wheat System of South Asia*. (Eds: J.K. Ladha, Olaf Erenstein, O. Erenstein and B. Hardy, Los Baños (Philippines): International Rice Research Institute), 279–296.
7. Rajender Parsad, Anshu Dixit, P. K. Malhotra, and V. K. Gupta (2007). Geoinformatics in precision farming: An overview. In *Geoinformatics Applications for*

Sustainable Development. New India Publishing Agency, New Delhi. (Eds: A.K. Singh and U.K. Chopra), 39-78.

8. V. K. Gupta and Rajender Parsad (2006). Statistical designing of experiments with emphasis on hill agriculture. In *Sustainable Production from Agricultural Watersheds in North West Himalayas*, Vivekananda Parvatiya Krishi Anusandhan Sansthan, Almora. (Eds: H.S. Gupta, A.K. Srivastava and J.C. Bhatt), 457-474.
9. B. S. Dhillon, M. Singh, Rajender Parsad, V. K. Gupta, and B. B. Singh (2004). Evaluation of plant genetic resources: Issues, advances and opportunities. In: Plant Genetic Resource Management. (Eds. B.S. Dhillon, R.K. Tyagi, A. Lal and S. Saxena), Narosa Publishing House, New Delhi, 242-265.
10. Rajender Parsad and V. K. Gupta (2004). Optimal designs for diallel cross experiments: An overview. In *Recent Advances in Mating Designs*. (Eds L. S. Kaushik and R. C. Hasija), Dhanpat Rai and Company (P) Ltd., New Delhi, 1.1-1.28.
11. Rajender Parsad and V. K. Gupta (2002). Optimal designs for diallel cross experiments: An overview. In *Recent Advances in Mating Designs*. (Eds L. S. Kaushik and R. C. Hasija), Dhanpat Rai and Company (P) Ltd., New Delhi. 3.1-3.23
12. Two blocks of Statistical Inference in the lecture notes on Probability and Statistics brought out by the Indira Gandhi National Open University, New Delhi, India.

16. Books' reviews written: 02

1. Written the book review of the "*Probability, Statistics, and Design of Experiments*" edited by Professor R.R. Bahadur, Wiley Eastern Limited Publication. The volume contains the proceedings of the International Symposium held in the honor of Late Professor R.C. Bose at the University of Delhi, Delhi 1988. The book review has appeared in *OPSEARCH, Journal of Operational Research Society of India*, 29, 1992.
2. Written the book review of "*Design and Analysis of Experiments*" by Angela Dean and Daniel Voss, a Springer Texts in Statistics, New York. The review has appeared in *Journal of Statistical Planning and Inference*, 87, 369–370, 2000.

17. List of technical/project reports published: 21

1. U. C. Sud, T. Ahmad, V. K. Gupta, H. Chandra, P. M. Sahoo, K. Aditya, M. Singh, and A. Biswas (2017). Methodology for estimation of crop area and crop yield under mixed and continuous cropping. FAO, Rome Publication 60, 4-87.
2. Sukanta Dash, Rajender Parsad, and V. K. Gupta (2014). Row-column designs for factorial experiments in two rows. IASRI, New Delhi. I.A.S.R.I./P.R.-11/2014.
3. Susheel Sarkar, Krishan Lal, and V. K. Gupta (2014). Main effects linear trend-free multi-level factorial experiments. IASRI, New Delhi. I.A.S.R.I./P.R.-04/2014.
4. B. N. Mandal, Rajender Parsad, and V. K. Gupta (2013). Application of optimization techniques for construction of incomplete block designs. IASRI, New Delhi. I.A.S.R.I./P.R.-01/2013.
5. L. M. Bhar and V. K. Gupta (2012). A study on multiple bio-assays. IASRI, New Delhi. IASRI/PR-08/2011.

6. Krishan Lal, Rajender Parsad, V. K. Gupta, and L. M. Bhar (2012). Analysis of experimental designs with t-Family of error distributions. IASRI, New Delhi. IASRI/PR-06/2012.
7. Aloke Lahiri, V. K. Gupta, A. Subba Rao, Y. Muralidharudu, Rajender Parsad, and Abhishek Rathore (2010). Planning, designing and analysis of experiments relating to AICRP on soil test crop response correlations. IASRI, New Delhi. I.A.S.R.I./P.R.-08/2010.
8. Krishan Lal, V. K. Gupta, P. K. Batra, and L. M. Bhar (2010). *Designs for Mixture Experiments in Agriculture*. ICAR-IASRI, New Delhi IASRI/PR-06/2010.
9. Seema Jaggi, V. K. Gupta, and Rajender Parsad (2008). Design and analysis of experiments for spatially correlated observations. IASRI, New Delhi. I.A.S.R.I./P.R.-02/2008
10. L. M. Bhar, Rajender Parsad, and V. K. Gupta (2008). Outliers in designed experiments. IASRI, New Delhi. I.A.S.R.I./P.R.-05/2008.
11. R. Srivastava, V. K. Gupta, and Rajender Parsad (2008). Study on robustness of block designs for biological assays against loss of data. IASRI, New Delhi. I.A.S.R.I./P.R.-09/2009.
12. Cini Varghese, A. R. Rao, V. K. Gupta, and Anu Sharma (2008). Statistical and algorithmic approach for improved estimation of treatment effects in repeated measurements designs. Project Report. Funded by Department of Science and Technology, New Delhi.
13. A. K. Nigam, Rajender Parsad, and V. K. Gupta (2006). Design and analysis of on-station and on-farm agricultural research experiments: A revisit. Joint Publication of IASDS, Lucknow and IASRI, New Delhi. I.A.S.R.I./P.R.-02/2006.
14. R. Srivastava, Rajender Parsad, and V. K. Gupta (2006). Studies on block designs for biological assays. IASRI, New Delhi. I.A.S.R.I./P.R.-05/2006.
15. Krishan Lal, Rajender Parsad and V. K. Gupta (2005). A study on trend free designs. IASRI, New Delhi. I.A.S.R.I./P.R.-02/2005.
16. Rajender Parsad, V. K. Gupta, P. K. Batra, R. Srivastava, Rajinder Kaur, Ajit Kaur, and Prawin Arya (2004). A diagnostic study of design and analysis of field experiments. IASRI, New Delhi. I.A.S.R.I./P.R.-01/2004.
17. V. K. Gupta, Rajender Parsad, Dipanker De, Hukam Chandra, and R.S. Singh (2001). Energy requirement in agricultural sector: A linear programming approach. Joint Publication of IASRI, New Delhi and CIAE, Bhopal. I.A.S.R.I./P.R.-21/2001.
18. Rajender Parsad, V. K. Gupta, and O. P. Khanduri (2000). Cataloguing and construction of variance balanced block designs: Computer algorithms for construction. IASRI, New Delhi. I.A.S.R.I./P.R.-16/2000.
19. R. Srivastava, V. K. Gupta, and Rajender Parsad (2000). Studies on optimality of block designs for making test treatments-control comparisons. IASRI, New Delhi. I.A.S.R.I./P.R.-12/2000.
20. Ashish Das, V. K. Gupta, and William I. Notz (1996). A class of universally optimal block designs. Technical Report No. 572, The Ohio Stat Univ., Columbus, USA.

21. Ashish Das, V. K. Gupta and Praggya Kisan (1990). E-optimal block designs under heteroscedastic model. Technical Report No. 1190, Stat.-Math. Division, ISI, Calcutta, India.

18. List of workshop proceedings published: 3

1. L. M. Bhar, Rajender Parsad, and V. K. Gupta (2008). Outliers in Designed Experiments. Workshop Series 3, IASRI, New Delhi.
2. R. Srivastava, Rajender Parsad, and V. K. Gupta (2005). Studies on Block Designs for Biological Assays. Workshop Series 2, IASRI, New Delhi.
3. A. K. Nigam, Rajender Parsad, and V. K. Gupta (2005). Design and Analysis of On-Station and On-Farm Agricultural Research Experiments: A Revisit. Workshop Series 0, IASRI, New Delhi.

19. List of special issues of journals brought out as guest editor: 11

1. Special Issue in the Fond Memories of C. R. Rao on "Life and Work of C. R. Rao (1920-2023): The Revolutionary of Statistical Sciences" (2024). Statistics and Applications, 22, 1-693. Guest Editors: B. K. Sinha, Bimal K. Sinha, S. D. Peddada, Thomas Mathew, Bhramar Mukherjee, Nairanjana Dasgupta, Syed Ejaj Ahmed, and V. K. Gupta
2. Special Issue "Contributions to Combinatorics, Design of Experiments, Linear Algebra and Related Areas" in the memory of Late Professor Aloke Dey (2021). Statistics and Applications, 19, 1-481. Guest Editors: V. K. Gupta, Sudhir Gupta, and Rajender Parsad
3. Special Issue "Challenges and Opportunities in Statistical Data Designing and Inference in the Emerging Global Scenario" in honour of twin statisticians Bimal and Bikas Sinha on their 75th Birthday (2020). Statistics and Applications, 18, 1-350. Guest Editors: V. K. Gupta, Saumyadipta Pyne, and Rajender Parsad
4. Special Volume to Felicitate Professor Arun Kumar Nigam on his 75th Birthday (2019). Statistics and Applications, 17, 1-280. Guest Editors: V. K. Gupta and Hukum Chandra
5. Special Issue on Survey Methodology and Applications (To felicitate Professor J. N. K. Rao on his 80th birthday) (2018). Statistics and Applications, 16, 1-409. Guest Editors: V. K. Gupta, Hukum Chandra, and Gauri S. Datta
6. Special Issue on "Statistical and Computational Methodologies for Massive Datasets" (2014). Journal of the Indian Society of Agricultural Statistics, 68, 117-321(Guest Editors V. K. Gupta, Naomi Altman, Snigdhansu Chatterjee, Kalyan Das, and Rajender Parsad).
7. Special Issue to "Felicitate Dr. Aloke Dey" (2009). Statistics and Applications, 6, 1-174 (Guest Editors V. K. Gupta, Rajender Parsad, and Ashish Das).
8. Special Issue "K. Kishen Memorial" (2008). Journal of the Indian Society of Agricultural Statistics, 62, 113-201 in honour of Late Dr. K. Kishen (Guest Editors V. K. Gupta, Rajender Parsad, and V. K. Bhatia).

9. Brought out the "Proceedings Volume of the International Conference on Statistics and Informatics in Agricultural Research", held at New Delhi in December 2006 to mark the Diamond Jubilee celebrations of the Indian Society of Agricultural Statistics. Journal of the Indian Society of Agricultural Statistics, (2007), 61, 95-279 (Guest Editors V. K. Gupta, V. K. Bhatia, Rajender Parsad, P. K. Malhotra, and H. V. L. Bathla).
10. As a Guest Editor brought out the Special Volume (Volume 3) in the honour of Dr. M. N. Das on his 80th birthday, Statistics and Application. This was released on the inaugural session of the 4th Annual Conference of the Society of Statistics, Computer and Applications, held at Rajkot during 24-26 November 2001.
11. As a member of the Editorial Committee, Brought out the Golden Jubilee Number 1996-97, Indian Society of Agricultural Statistics. This special issue of the Journal was released during the inaugural session of the Golden Jubilee Conference held at Vigyan Bhawan. The Editorial Committee for this comprised of A. Dey, V. K. Gupta, V. K. Bhatia, and A. K. Srivastava.

Besides this 09 Special Proceedings of the Conferences 19-27, SSCA (Guest Editor)

20. Copyrights obtained: 08

Registrar Copyright Office, New Delhi has granted the following copyrights in which VKG is an author / co-author:

1. *Statistical Package for Factorial Experiments (SPFE 1.0)* on June 06, 2013. Copyright No. SW-6674/2013: Sangeeta Ahuja, Rajender Parsad, V. K. Gupta, and P. K. Malhotra.
2. *Statistical Package for Augmented Designs (SPAD)* on June 06, 2013. Copyright No. SW-6675/2013: Abhishek Rathore, Rajender Parsad and V. K. Gupta.
3. Design Resources Server on June 07, 2013. Copyright No. L-46452/2013: Rajender Parsad and V. K. Gupta.
4. *Monograph on Hadamard Matrices* on July 04, 2013. Copyright No. L-51175/2013: V. K. Gupta, A. Dhandapani and Rajender Parsad.
5. *Monograph α-designs* on July 24, 2013. Copyright No. L-51734/2013: Rajender Parsad, V. K. Gupta, P. K. Batra, S. K. Satpati, and Pabitra Biswas.
6. *Statistical Analysis of Agricultural Experiments Part 1: Single Factor Experiments.* Copyright No. L-67363/2017: B. N. Mandal, L. M. Bhar, Rajender Parsad, and V. K. Gupta on July 17, 2017.
7. *Significance of Experimental Designs in Agricultural Research.* Copyright No. L-73008/2018: B. N. Mandal, Rajender Parsad, V. K. Gupta on February 02, 2018.
8. *Sample Survey Resources Server.* Copyright No. SW-9073/2017: Anshu Bharadwaj, Hukum Chandra, Rajender Parsad, U. C. Sud, V. K. Gupta on May 18, 2017.

Acknowledgements

Sincerest gratitude to the SSCA/Concerned Committee headed by Dr. D. V. V. Ramana, a valued colleague, for recommending my name for V. K. Gupta Endowment Award for Achievements in Statistical Thinking and Practice Lecture in the name of my friend, philosopher and guide Vinod Kumar Gupta. I bow to my teachers and mentors who have guided me throughout my life. I would also like to place on record my sincerest thanks to

गुरुर्ब्रह्मा गुरुर्विष्णुर्गुरुर्देवो महेश्वरः ।
गुरुरेव परं ब्रह्म तस्मै श्रीगुरुवे नमः ॥

the anonymous reviewer for valuable suggestions that have helped in better presentation of the manuscript.

Conflict of interest

The author does not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Avadhani, M. S. and Sukhatme, B. V. (1973). Controlled sampling with equal probabilities and without replacement. *International Statistical Review*, **41**, 175-182.

Chakrabarti, M. C. (1963). On the use of incidence matrices in sampling from finite populations. *Journal of the Indian Statistical Association*, **1**, 78-85.

Gopinath P. P., Parsad, R., and Mandal, B. N. (2018). Two-dimensional balanced sampling plans excluding adjacent units under sharing a border and island adjacency schemes. *Communications in Statistics - Theory and Methods*, **47**, 712-720.

Gurney, M. and Jewett, R. S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, **70**, 819–821.

Hedayat, A. S., Rao, C. R., and Stufken, J. (1988). Sampling plans excluding contiguous units. *Journal of Statistical Planning and Inference*, **19**, 159-170.

Magnussen, S. (2009). A balanced repeated replication estimator of sampling variance for apparent and predicted species richness. *Forest Science*, **55**, 189-200.

McCarthy, P. J. (1966). Pseudo-replication: Half samples. *Review of International Statistical Institute*, **37**, 239-264.

Rao, J. N. K. and Shao, J. (1996). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Association*, **91**, 343-348.

Rao, J. N. K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, **86**, 403-415.

Saigo, H., Shao, J., and Sitter, R.R. (2001). A repeated half sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology*, **27**, 189-196.

Stufken, J., Song, S. Y., See, K., and Driessel, K. R. (1999). Polygonal designs: some existence and non-existence results. *Journal of Statistical Planning and Inference*, **77**, 155-166.

Wu, C. F. J.(1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, **78**, 181-188.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 63-74

Enabling Artificial Intelligence in Health: Perspectives on Health Data Challenges, Mitigation Strategies, and the Current Ecosystem in India

Vishal Deo and Kratika Pandiya

National Institute for Research in Digital Health
Indian Council of Medical Research, New Delhi-110023, India

Received: 18 July 2026; Revised: 8 August 2026; Accepted: 10 August 2026

Abstract

Artificial intelligence (AI) and machine learning (ML) are increasingly embedded in health systems for clinical decision support, disease surveillance and health system planning, yet their performance and fairness depend fundamentally on the quality and governance of underlying data. This paper examines health data challenges, mitigation strategies for these challenges, and explore the current health data ecosystem in India.

First, the paper synthesises key challenges in health data sharing and quality, including limited willingness to share, privacy and ownership concerns, institutional silos, irregular interoperability, and problems of accuracy, completeness, timeliness and representativeness. It also discusses the importance of statistical methods in mitigating data quality issues. It highlights how adoption of FAIR (Findable, Accessible, Interoperable, Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, Technology) principles can underpin trustworthy repositories and data pipelines.

Second, it analyses data-related sources of bias in predictive models, distinguishing data-induced algorithmic bias from two forms of training-data bias: Type I, arising from selection and under-representation of specific subgroups, and Type II, arising from explicit use of social constructs such as race or caste.

Third, the paper maps India's health and demographic data ecosystem, including national surveys, surveillance networks, health system data, and research data repositories, and discusses opportunities and limitations for AI training and validation. Finally, it reviews national digital health frameworks, notably the Ayushman Bharat Digital Mission (ABDM), as foundational infrastructure for interoperable, consent-mediated electronic health records.

Key words: Health data infrastructure; Data repositories; Data quality; AI in health; Data sharing principles; Digital health; Data bias.

Corresponding Author: Vishal Deo
Email: vishaaldeo@gmail.com

1. Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly promoted as transformative tools for clinical decision support, disease surveillance, and health system planning. Rapid growth in computing power, from teraflop to petaflop-scale hardware, and the availability of large digital datasets have made it feasible to deploy complex predictive models in routine healthcare settings (Zhu *et al.* (2023), Hassani and MacFeely (2023)).

However, the performance and fairness of these models ultimately depend on the quality, representativeness and governance of the data used to train and evaluate them. While digital health data in India are expanding, they remain fragmented, uneven in quality and often inaccessible to researchers (Salvatore *et al.* (2026), Maity *et al.* (2024), and Hassani and MacFeely (2023)). Statistics and data science provide the conceptual and methodological foundation for addressing these challenges, from designing representative data collection systems and quantifying bias, to developing robust validation frameworks and enabling responsible data sharing.

This paper provides a synthesized review of the challenges of data sharing and data quality, data induced biases in AI/ML algorithms, the role of statistical methods in mitigating these biases, the current health and demographic data ecosystem in India, and national frameworks and support system for Digital Health Data/ Electronic Health Record (EHRs).

2. Health data challenges

Health data for AI/ML in healthcare originate from multiple sources: EHRs, hospital management information system (HMIS), disease surveillance platforms, national surveys, registries and increasingly from patient-generated data streams. Despite this apparent abundance, several structural challenges constrain the effective use of these data for algorithm development and evaluation, particularly in the Indian context (Salvatore *et al.* (2026) and Maity *et al.* (2024)).

Two broad domains of concern are: (i) limitations and barriers in data sharing, and (ii) persistent problems in data quality. These domains are closely coupled, since poor quality and lack of standardisation amplify privacy, governance and trust concerns, while restricted access limits the ability of independent researchers to innovate, and also to diagnose and correct biases in deployed AI systems.

2.1. Data sharing

Lack of willingness to share data is a primary barrier, driven by concerns about privacy and confidentiality, unclear data ownership, and institutional silos (Baines *et al.* (2024) and Van Panhuis *et al.* (2014)). Many health facilities and programmes treat their data as proprietary assets, which, combined with weak data governance frameworks, leads to highly restricted access even for bona fide research purposes.

Interoperability challenges further inhibit data sharing. When health information systems use incompatible formats, coding standards and identifiers, aggregating data across platforms becomes technically difficult and costly (Van Panhuis *et al.* (2014)). Security concerns, limited infrastructure and absence of standardised data sharing pipelines – for

example, secure APIs, de-identification services and governance contracts – mean that even when data exist in digital form, they seldom move seamlessly between providers, programmes and researchers.

Wilkinson *et al.* (2016) introduced the Findable, Accessible, Interoperable, Reusable (FAIR) principles for data sharing. The aim of FAIR principles is to provide practical guidance for anyone seeking to make their data more reusable. The FAIR principles explicitly prioritise making data automatically discoverable and usable by machines, while still supporting reuse by individual users (Martínez-García *et al.* (2023)). Lin *et al.* (2020) underscores the importance of adopting Transparency, Responsibility, User focus, Sustainability, Technology (TRUST) principles for health data repositories. Applying FAIR/TRUST in practice implies building trustworthy digital repositories with clear organisational mandates, sustainable funding, robust technical infrastructure, external certification (such as CoreTrustSeal or ISO 16363), and transparent policies for access, consent and reuse.

2.2. Data quality

Data quality issues manifest along several dimensions, including accuracy, completeness, consistency, timeliness, representativeness, validity and source credibility (Schmidt *et al.* (2021)). In many Indian programmes, routine data are collected primarily for service reporting or administrative purposes, rather than for analytic robustness, which can lead to under-reporting, misclassification and variable adherence to definitions across states and facilities (Blanchard *et al.* (2020)).

Representativeness is particularly critical for AI applications. If datasets over-represent certain regions, facility types or socio-demographic groups, the resulting models may perform poorly, or even harmfully, for under-represented populations (Hassani and MacFeely (2023)). Timeliness and interoperability also affect the feasibility of using health data for real-time surveillance or adaptive decision support, since lagged or siloed datasets lead to outdated or incomplete signals.

Data labelling and annotation quality, and granularity are other important aspects of data quality, especially for imaging and sensor data. Without standardised annotation protocols, multi-reader verification and detailed metadata on devices, acquisition conditions and grading procedures, AI models trained on such datasets can exhibit unpredictable performance when deployed in different settings. Squires *et al.* (2023) quantifies how variability in radiologist breast-density labels affects deep-learning performance, demonstrating that inconsistent grading leads to unpredictable model behaviour and reduced reliability when deployed. Multusch *et al.* (2025) demonstrated that annotation quality varies with reader expertise and that noisy labels from less experienced annotators degrade AI performance, underscoring the need for structured multi reader verification and quality control in radiology datasets.

3. Data induced bias in AI algorithms

Bias in medical AI can be defined as systematic error or undesired dependence of an algorithm on specific attributes, often correlates of demographic or socio-economic groups, that leads to unequal performance or outcomes. Recent reviews have catalogued multiple

forms of bias in AI-based models for medical applications, highlighting implications for clinical decision-making, triage and resource allocation (Cross *et al.* (2024) and Mittermaier *et al.* (2023)).

Data-induced algorithmic bias occurs when patterns present in training data, including measurement errors, missingness, confounding and social inequities, are inadvertently learned and amplified by the model. Bias is not only about explicit encoding of protected attributes, but about any undesired dependence on data features that reflect inequitable structures in healthcare access, financing or outcomes. For example, if historical data record lower utilisation of specialist services among marginalised communities due to barriers in access, an AI system trained to predict future utilisation may learn that these groups "need" fewer referrals, thereby perpetuating under-treatment. Similarly, if cost data reflect differential spending patterns rather than clinical need, models that use cost as a proxy outcome can embed socio-economic bias in risk scores and care prioritisation.

For the sake of discussion, we categorize data-related sources of bias into two broad categories: (i) training data bias Type I, arising from selection and representativeness; and (ii) training data bias Type II, arising from the use of social constructs such as race, ethnicity or caste as explicit variables in algorithms.

3.1. Training data bias Type-I: selection bias/ imbalanced sample

Here our discussion corresponds to selection bias and imbalanced sample sizes, *i.e.*, over or under-representation of particular subpopulations in the data used to train AI models. This leads to worse performance and systematic underestimation of risk for under-represented groups.

A widely discussed example is the algorithm used by US hospitals to prioritise patients for extra care, analysed by Obermeyer *et al.* (2019). The algorithm did not include race as an explicit variable, but used healthcare cost history as a proxy for healthcare needs. Because Black patients historically incurred lower healthcare costs despite similar or greater need, the algorithm assigned them lower risk scores than white patients, resulting in fewer referrals for additional care at equivalent levels of morbidity.

A validation and implementation study for real world evaluation of AI-driven diabetic retinopathy (DR) screening in India done by Duggal *et al.* (2025) reported that AI algorithms validated on international datasets showed substantially degraded performance when deployed in Indian public screening sites. Sensitivity dropped to 60–80% and specificity to as low as 14% on Indian images, largely due to the absence of standardised, high-quality and regionally representative training data and poor documentation of device types, image quality and grading procedures.

Importantly, AI developers were often unable or unwilling to specify the provenance, characteristics or size of their training datasets, and metadata on camera type, image quality, population demographics and labelling protocols were rarely accessible (Duggal *et al.* (2025)). Such opacity prevents meaningful assessment of contextual validity and bias, and underscores the need for transparent data curation standards and regulatory requirements for documentation.

3.2. Training data bias Type-II: unintended bias due to use of social constructs as factors

These biases arise when social constructs such as race, ethnicity or caste are directly encoded in clinical algorithms, often without robust evidence on their biological or public health relevance. These constructs are not biological categories and may introduce unintended bias if used naively in risk prediction or diagnostic models.

A key example is the use of race in estimating kidney function, specifically estimated glomerular filtration rate (eGFR). Mohottige *et al.* (2023) reported limited justification for using "Black race" as an adjustment factor in eGFR equations. They showed how including race led to overestimation of kidney function in Black patients, delaying referral for nephrology care and transplantation, thereby exacerbating health inequities.

In such situations, mitigation strategies include replacing race-based adjustments with variables that more directly capture socio-economic deprivation, such as area-level indicators derived from zip codes or location information (Cross *et al.* (2024)). More broadly, the message is that algorithm designers should critically interrogate any variable that reflects historical discrimination or structural inequity, and preferentially model underlying mechanisms (for example, access barriers, environmental exposures, co-morbidities) rather than coarse social categories.

Other training-data-based biases include non-random missingness (for instance, laboratory data available only for those who can afford tests), lack of capture of important clinical variables in routine datasets, and misclassification or noisy labels in electronic records and registries.

3.3. Role of statistical methods in mitigating these biases

From a statistical perspective, handling such biases requires explicit modelling of missing data mechanisms, careful design of validation studies, and robust sensitivity analyses that assess the impact of plausible bias scenarios on model performance and fairness (Hauschildt *et al.* (2025), Cross *et al.* (2024) and Mittermaier *et al.* (2023)). Selection bias and imbalanced samples can be corrected using inverse-probability weighting, propensity-score matching, and resampling methods such as Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and stratified under/oversampling (Krautenbacher *et al.* (2017) and Chawla *et al.* (2002)). Under- or over-representation is addressed through data rebalancing, reweighting, and fairness-aware pre-, and post-processing pipelines, covering fairness constraints, adversarial debiasing, and threshold calibration (Chen *et al.* (2024)). Non-random missingness requires multiple imputation, inverse-probability weighting, or selection/pattern-mixture models instead of complete-case analysis, since ignoring the missingness mechanism biases estimates (Little and Rubin (2019)). Misclassification is handled through inverse-propensity reweighting of mislabeled points and regression-calibration-type corrections, improving classifier reliability when trained on biased or noisy data (Song *et al.* (2025)).

4. Current health and demographic data ecosystem in India

India has developed a rich, though fragmented, health and demographic data ecosystem comprising large-scale household surveys, disease surveillance platforms, civil registration systems and facility-based HMIS. In this article, we have reviewed key components of this ecosystem and their relevance for AI-enabled health applications, while also noting limitations in coverage, granularity and accessibility.

From an AI/ML standpoint, these data sources have the potential to provide essential ground truth and contextual information for training, validating and monitoring predictive models related to maternal and child health, infectious diseases, non-communicable diseases and health service utilisation. Yet, their differential periodicity, representativeness and openness to researchers pose challenges for integrating them into unified data strategies.

4.1. Survey datasets

The National Family Health Survey (NFHS) is one of the largest demographic and health surveys globally, conducted by the Ministry of Health and Family Welfare (MoHFW) with International Institute for Population Sciences (IIPS), Mumbai, as the nodal agency (IIPS (2026)). Six rounds have been completed between 1992-93 and 2023-24. Since NFHS-4, the survey provides district-level estimates on fertility, mortality, maternal and child health, family planning, nutrition, and selected non-communicable disease indicators.

Other survey datasets include the District Level Household and Facility Survey (DLHS) and Annual Health Survey (AHS), which historically provided District Level reproductive and child health and mortality indicators in selected states, as well as health-related rounds of the National Sample Survey (NSS) on morbidity, health expenditure, disability, and aging. Although DLHS and AHS have been discontinued, their functions have largely been absorbed into NFHS and HMIS while NSS continues to provide critical information on out-of-pocket expenditure, catastrophic spending and disability that can inform equity-focused AI models Dandona *et al.* (2016)).

Limitations of such survey datasets include high cost and resource intensity, recall and self-reporting biases, limited coverage of adult health and mental health, and challenges in comparability over time due to changes in questionnaires and definitions. Moreover, their periodic nature means that rapidly evolving health issues, such as emerging infectious diseases or changing risk factors, are not captured in real time, which restricts their direct use for dynamic model updating.

4.2. Disease surveillance datasets

Disease surveillance platforms provide near real-time information on epidemic-prone conditions and outbreaks. The Integrated Disease Surveillance Programme (IDSP), launched in the early 2000s and now part of the National Health Mission, aims to strengthen decentralised, laboratory-based IT-enabled surveillance for epidemic-prone diseases and to support early detection and response through trained rapid response teams
(Ministry of Health and Family Welfare (2026)).

Building on IDSP, the Integrated Health Information Platform (IHIP) is a web-based extension launched in 2019 that enables near real-time reporting, analysis and geo-tagged facility data for over 50 conditions (https://ihiplearning.in/idsp/). In parallel, the Virus Research and Diagnostic Laboratory (VRDL) network coordinated by the Department of Health Research (DHR) and the Indian Council of Medical Research (ICMR) has established more than 160 laboratories that generate test data on viral diseases, shared in near real time with IDSP and other programmes (Department of Health Research (2026)).

From an AI perspective, these platforms offer high-frequency signals for outbreak detection, syndromic surveillance and dynamic risk prediction, but there are some limitations as well. Facility-based surveillance may not be representative of community-level disease burden; fragmentation across systems leads to duplication and inconsistency; private sector facilities, which deliver a substantial share of care in India, often do not report systematically; and data are rarely open, timely or user-friendly for external researchers.

These limitations constrain the ability to train robust forecasting models or evaluate algorithm performance across heterogeneous settings. Addressing them will require stronger mandates and incentives for reporting, integration of private sector data, improved data quality assurance, and governance frameworks that support controlled access for research and innovation.

4.3. Hospital management information system

HMIS of the MoHFW is a nationwide facility-based reporting platform covering sub-centres, primary health centres, community health centres, district hospitals and higher facilities (https://www.hmis.mohfw.gov.in). Launched in 2008, HMIS provides monthly aggregated indicators on service utilisation, maternal and child health, immunisation, family planning, non-communicable diseases and programme performance.

HMIS, together with NFHS, now serves as a primary source of district-level health and demographic data in India. Yet, HMIS faces challenges of limited private sector participation, variable timeliness and completeness across states, frequent staff turnover among data entry personnel, and inconsistent validation and feedback mechanisms (Gupta *et al.* (2024)).

For AI/ML applications, HMIS offers a potential backbone for learning models of service utilisation, coverage gaps and programme performance at fine geographic granularity. However, realising these potential demands investments in data quality, standardisation of coding and aggregation rules, and creation of anonymised research-ready data extracts. Linking HMIS with EHRs and registries could eventually support more granular, patient-level analytics within robust governance and consent frameworks.

4.4. National repositories of health and biomedical research datasets

As a new initiative, ICMR has established the National Health Research Data Repository (https://data.icmr.org.in/) with the vision to empower researchers with secure, centralized access to high-quality datasets, advancing medical research innovation while safeguarding data integrity and privacy through rigorous verification standards. High-quality datasets, along with comprehensive metadata, data snapshots, and data descriptions are available on the portal which can be seen to anyone through free portal registration. Cur-

rently, researchers are provided access to complete datasets for analysis through the federated principle, whereby analysis code is sent to the data rather than the data being shared with researchers.

Another important initiative in this direction is the Indian Biological Data Centre (IBDC) established by the Department of Biotechnology (https://ibdc.dbt.gov.in). It is the first national repository for life science data in India. IBDC is mandated to archive all life science data generated from publicly funded research in India. As a national life-sciences data archive IBDC houses sub-repositories like the Indian Nucleotide Data Archive, Indian Proteome Databank, Indian Structural Data Archive, Indian Biomedical Image Archive, and Indian Phenome Repository.

As a more generic repository, AIKosh (https://aikosh.indiaai.gov.in/home) is the national AI datasets and resources platform launched under the IndiaAI Mission of the Ministry of Electronics and Information Technology (MeitY). It is designed to democratize access to high-quality, India-specific AI artefacts to researchers and innovators. It serves as a unified hub offering over 14,000 datasets, over 300 AI models, and more than 30 toolkits spanning sectors such as health, agriculture, education, and governance, along with benchmarks, use cases, and a secure sandbox environment for model training and experimentation. Organizations, including academic institutions, startups, and non-profits, can contribute non-personal, anonymized datasets and models via an Expression of Interest process, retaining control over licensing while ensuring compliance with ethical AI and data-governance standards.

5. Overview of available frameworks and support system for digital health data/ EHRs

Recognising the need for interoperable, citizen-centric digital health infrastructure, India has launched several national initiatives that define standards, registries and governance mechanisms for health data and EHRs. Two key components of this ecosystem are the Ayushman Bharat Digital Mission (ABDM) and the National Resource Centre for EHR Standards (NRCeS).

5.1. Ayushman Bharat Digital Mission (ABDM)

ABDM aims to create a national digital health ecosystem that supports universal health coverage in an efficient, accessible, inclusive, affordable, timely and safe manner. It is designed as the backbone for integrated digital health infrastructure, leveraging India's broader public digital rails, including Aadhaar, Unified Payments Interface and widespread mobile connectivity, while adhering to the guiding principles of the National Digital Health Blueprint (https://abdm.gov.in).

Key architectural features of the ABDM include decentralised data storage, wherein health data and applications are held at points of care; patient control over processing of their health data; and consent-driven, link-based data access, with entities able to access records only through secure links subject to permissions and consent (National Health Authority (2022)). The Ayushman Bharat Health Account (ABHA) number serves as a unique digital identifier that threads longitudinal health records across facilities, with the individual

retaining control over sharing and linkage.

ABDM also maintains two major registries: the Health Facility Registry (HFR), a comprehensive repository of public and private health facilities across systems of medicine, and the Healthcare Professionals Registry (HPR), which records registered and verified healthcare practitioners. As of July 2026, government reports indicate over 94 crores of ABHA accounts created, over 106 crores health records linked, and more than five lakh facilities and 10 lakh professionals onboarded, underscoring the rapid expansion of India's digital health infrastructure (https://dashboard.abdm.gov.in/abdm/).

From a statistical and AI standpoint, ABDM introduces both opportunities and responsibilities. Robust, consent-mediated linkage of patient-level records across providers can enable more accurate risk prediction, longitudinal analyses and personalised decision support. However, realising these gains requires strict adherence to privacy-preserving data management policies as per India's Digital Personal Data Protection (DPDP) Act, 2023, clear rules against re-identification, and mechanisms for independent algorithmic auditing anchored in transparent data documentation.

5.2. National Resource Centre for EHR Standards (NRCeS)

To accelerate adoption of interoperable EHR standards, the MoHFW released national EHR standards in 2013 and updated them in 2016, and established the NRCeS at Centre for Development of Advanced Computing (C-DAC), Pune.

NRCeS functions as a centre of excellence and single point of contact for assistance in developing, implementing and using EHR standards in India, providing a knowledge base and tools that facilitate adherence to these standards (https://www.nrces.in/).

The standards cover data formats, coding systems (such as SNOMED CT and ICD), interoperability specifications (such as FHIR/HL7) and security requirements necessary for exchange of health information across systems. For AI developers, compliance with these standards is critical. It ensures that models trained on EHR data are portable across platforms, that variables are consistently defined and coded, and that model outputs can be integrated into clinical workflows without compromising data integrity or privacy. In combination, ABDM and NRCeS provide the scaffolding for a more coherent digital health data ecosystem in India.

6. Conclusion

A review of the current data ecosystem of India shows that India possesses a diverse array of health and demographic data sources – from NFHS, NSS and legacy district surveys to IDSP/IHIP, VRDL, HMIS, ABDM registries and emerging EHR implementations. Yet these assets remain fragmented across institutions and programmes, with heterogeneous standards, variable quality and limited interoperability. As a result, researchers and innovators often struggle to obtain timely, well-documented data needed to develop, validate and monitor AI/ML models that are both accurate and equitable.

Bridging this gap requires concerted efforts on multiple fronts. First, policymakers and programme managers must invest in harmonising data standards, strengthening qual-

ity assurance, and enabling secure, consent-mediated data sharing pipelines across surveys, surveillance systems and facility-based platforms. Second, health statisticians and data scientists should embed rigorous bias assessment, missing data modelling and fairness metrics into the lifecycle of AI system development, drawing on emerging literature on bias in medical AI and leveraging India-specific validation studies such as the diabetic retinopathy screening work highlighted in the article. Third, researchers themselves need to change behaviour and culture around data sharing. While ethical and legal safeguards are indispensable, reflexive non-disclosure and siloed data ownership impede collective learning and stifle innovation. Embracing FAIR and TRUST principles, contributing anonymised datasets and code to well-governed repositories, and collaborating across disciplines can help create a virtuous cycle where better data enable better models, which in turn inform better policies and clinical decisions.

Ultimately, enabling AI in health is not merely a technical endeavour. It is a statistical and governance challenge that hinges on how societies collect, curate, interpret and share health data. For India, the next decade offers an opportunity to transform its rich but fragmented data landscape into an integrated, interoperable and researcher-friendly ecosystem that supports responsible AI innovation and advances health equity.

Acknowledgements

The authors are grateful to the reviewers for their valuable comments and suggestions, and the editors for their guidance and counsel to greatly imrpove the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Baines, R., Stevens, S., Austin, D., Anil, K., Bradwell, H., Cooper, L., Maramba, I., Chatterjee, A., and Leigh, S. (2024). Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. *Journal of Medical Internet Research*, **26**, e50421.

Blanchard, J., Washington, R., Becker, M., Vasanthakumar, N., Gopal, K., and Sarwal, R. (2020). Vision 2035–public health surveillance in India. A White Paper.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Chen, F., Wang, L., Hong, J., Jiang, J., and Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, **31**, 1172–1183.

Cross, J. L., Choma, M. A., and Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, **3**, e0000651.

Dandona, R., Pandey, A., and Dandona, L. (2016). A review of national health surveys in India. *Bulletin of the World Health Organization*, **94**, 286.

Duggal, M., Chauhan, A., Gupta, V., Kankaria, A., Budhija, D., Verma, P andMiglani, V., Syal, P andKaur, G., Kumar, L., Mutyala, N., Bezbaruah, R., Sood, N., Kernohan, A., Menon, G., and Vale, L. (2025). Real-world evaluation of AI-driven diabetic retinopathy screening in public health settings: Validation and implementation study. *JMIR Medical Informatics*, **13**, e67529.

Gupta, A., Sharma, A., and Jha, D. K. (2024). Overcoming obstacles step by step: A comprehensive review of challenges and strategies in implementing hospital management information systems in India.

Hassani, H. and MacFeely, S. (2023). Driving excellence in official statistics: Unleashing the potential of comprehensive digital data governance. *Big Data and Cognitive Computing*, **7**, 134.

Hauschildt, K. E., Pan, A., Bernstein, T., Admon, A. J., Mukherjee, B., Iwashyna, T. J., and Rountree, L. (2025). Consideration of sociodemographics in machine learning-driven sepsis risk prediction. *Critical Care Medicine*, **53**, e1815–e1820.

IIPS (2026). National family health survey user portal. retrieved July 17, 2026, from https://www.nfhsiips.in/nfhsuser/index.php.

Krautenbacher, N., Theis, F. J., and Fuchs, C. (2017). Correcting classifiers for sample selection bias in two-phase case-control studies. *Computational and Mathematical Methods in Medicine*, **2017**, 7847531.

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ..., and Westbrook, J. (2020). The trust principles for digital repositories. *Scientific Data*, **7**, 144.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Maity, D., Satish, R., Jadeja, D. A., Dharmaraju, R., Chandru, V., Sundaresan, R., Singh, H., and Pal, D. (2024). MIDAS: a new platform for quality-graded health data for AI-enabled healthcare in India. *Nature Medicine*, **30**, 2704–2705.

Martínez-García, A., Alvarez-Romero, C., Román-Villarán, E., Bernabeu-Wittel, M., and Parra-Calderón, C. L. (2023). FAIR principles to improve the impact on health research management outcomes. *Heliyon*, **9**.

Mittermaier, M., Raza, M. M., and Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *NPJ Digital Medicine*, **6**, 113.

Mohottige, D., Olabisi, O., and Boulware, L. E. (2023). Use of race in kidney function estimation: Lessons learned and the path toward health justice. *Annual Review of Medicine*, **74**, 385–400.

Multusch, M. M., Hansen, L., Heinrich, M. P., Berkel, L., Saalbach, A., Schulz, H., Wegner, F., Barkhausen, J., and Sieren, M. M. (2025). Impact of radiologist experience on AI annotation quality in chest radiographs: a comparative analysis. *Diagnostics*, **15**, 777.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, **366**, 447–453.

Salvatore, M., Wang, Y., Syal, P., Wahl, B., Parikh, H., Deo, V., ..., and Mukherjee, B. (2026). The case for an integrated biobanking initiative in South Asia. *The Lancet Regional Health-Southeast Asia*, **49**.

Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmidt, B., Sauerbrei, W., and Richter, A. (2021). Facilitating harmonized data quality assessments. a data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, **21**, 63.

Song, B., Zhao, S., Dang, L., Wang, H., and Xu, L. (2025). A survey on learning from data with label noise via deep neural networks. *Systems Science & Control Engineering*, **13**, 2488120.

Squires, S., Harkness, E. F., Evans, D. G., and Astley, S. M. (2023). The effect of variable labels on deep learning models trained to predict breast density. *Biomedical Physics & Engineering Express*, **9**, 035030.

Department of Health Research (2026). Setting up of nation-wide network of laboratories for managing epidemics and natural calamities (viral research and diagnostic labs). ministry of health and family welfare, government of India. retrieved July 18, 2026, from https://www.dhr.gov.in/offerings?page=1.

Ministry of Health and Family Welfare (2026). History of the integrated disease surveillance programme (idsp). retrieved July 18, 2026, from https://ihiplearning.in/idsp/#/about/history.

National Health Authority (2022). A brief guide on ayushman bharat digital mission and its various building blocks. accessed july 2026 from `https://abdm.gov.in/strapicms/uploads/ABDM_Handbook_19_10_2022_24c5078481.pdf`.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., and Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health*, **14**, 1144.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ..., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**, 1–9.

Zhu, S., Yu, T., Xu, T., Chen, H., Dustdar, S., Gigan, S., ..., and Pan, Y. (2023). Intelligent computing: the latest advances, challenges, and future. *Intelligent Computing*, **2**, 0006.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 75-95

Waiting Time Until Success

Jyotirmoy Sarkar

Department of Mathematical Sciences
Indiana University Indianapolis, IN 46202, USA

Received: 30 January 2026; Revised: 04 August 2026; Accepted: 19 August 2026

Dedication

This research is respectfully dedicated to honor my guruji, mamababu, and guide, Professor Bikas Kumar Sinha, whose association, blessings, and collaboration can reduce the waiting time until success for any student of probability and statistics.

Abstract

Waiting time until success is a fundamental concept in probability, with the geometric distribution providing the classic example. We investigate waiting times associated with more elaborate notions of success, including prescribed patterns of outcomes, events in multinomial experiments, features of finite Markov chains, and return times of random walks on graphs. A common theme throughout the paper is the use of renewal-type recursions derived from one-step transition probabilities. They generate the probability mass functions of the waiting times through matrix iterations. The unified methodology produces explicit distributions for a variety of stochastic processes and reveals connections between pattern-occurrence problems, Markov chains, and random walks.

Key words: Probability mass function; Lack of memory; Renewal time; Recursive relation; Matrix inversion; Stationary distribution.

AMS Subject Classifications: 97K60, 60J10

1. Introduction

Suppose that we toss a biased coin with a known $P(\text{Heads}) = p \in (0,1)$. Then the number X of tosses until the first head appears is said to have the geometric(p) distribution. Its probability mass function (PMF) is given by

$$p_x = P\{X = x\} = q^{x-1}p, \quad \text{for } x = 1, 2, 3, \ldots, \tag{1}$$

where $q = 1 - p = P(\text{Tails})$. The probability masses $p_1 = p, p_2 = qp, p_3 = q^2p, \ldots$ of X form a geometric series. Hence, X is called a geometric random variable (RV).

Corresponding Author: Jyotirmoy Sarkar
Email: jsarkar@iu.edu

Most textbooks (see, for example, Devore and Berk (2007), Dudewicz and Mishra (1988), Hoel *et al.* (1971), Rohatgi (2003)) develop the properties of a geometric RV starting from the PMF (1). We list a few such properties below.

(G1) Total probability: The sequence of probability masses is geometric, and adds to

$$p + qp + q^2p + \ldots = p[1 + q + q^2 + \ldots] = p(1-q)^{-1} = 1.$$

No mass escapes to infinity; the process stops in finite time with probability one.

(G2) Probability generating function (PGF): For $0 < t < q^{-1}$,

$$P_X(t) = E[t^X] = pt + qpt^2 + q^2pt^3 + \ldots = pt\ [1 + (qt) + (qt)^2 + \ldots] = pt(1-qt)^{-1}.$$

(G3) First moment: The expected waiting time until the first head appears is

$$\begin{aligned} E[X] &= p + 2qp + 3q^2p + \ldots = p\ \frac{d}{dq}\{q + q^2 + q^3 + \ldots\} \\ &= p\ \frac{d}{dq}\{1 + q + q^2 + q^3 + \ldots\} = p\ \frac{d}{dq}(1-q)^{-1} = p(1-q)^{-2} = p^{-1}. \end{aligned}$$

(G4) Moment generating function (MGF): For $0 < t < -\ln q$,

$$\begin{aligned} M_X(t) = E[e^{tX}] &= pe^t + qpe^{2t} + q^2pe^{3t} + \ldots \\ &= pe^t\ [1 + (qe^t) + (qe^t)^2 + \ldots] = pe^t(1-qe^t)^{-1}. \end{aligned}$$

(G5) Lack of memory property: For $x \geq 1$, we have $P\{X > x\} = q^xp + q^{x+1}p + \ldots = q^x$. Consequently, for $x, y \geq 1$, we have

$$P\{X > x + y | X > x\} = \frac{P\{X > x + y\}}{P\{X > x\}} = \frac{q^{x+y}}{q^x} = q^y = P\{X > y\}.$$

In the above, properties G1–G5 are proved by evaluating a geometric series. Below, we derive these properties starting from the **renewal property** given by

$$X = \begin{cases} 1 & \text{with probability } p \\ 1 + \tilde{X} & \text{with probability } q, \end{cases} \tag{2}$$

where $\tilde{X}$ has the same probability distribution as X. In other words, after the first toss, either a success (a head) has occurred and the process stops, or the process has *renewed* itself in the sense that the number of additional tosses (after the first toss that did not yield a success) until a success occurs has the same distribution as the original RV X itself. Equivalently, (2) can be written as

$$p_x = \begin{cases} p & \text{if } x = 1 \\ q\,p_{x-1} & \text{if } x > 1. \end{cases} \tag{3}$$

Writing all probability masses as a vector $\mathbf{p}_X = (p_1, p_2, \ldots)$, (3) implies $\mathbf{p}_X = (p, q\,\mathbf{p}_X)$. Thereafter, by mathematical induction on x, one can prove that (3) implies (1). Below, we show that properties G1–G5 follow from (3), without requiring either evaluation of a geometric series or a differentiation technique. We prove these relatively easy results as a precursor to the extensions we have in mind.

(G1) Total probability: Writing $T = \sum_{x=1}^{\infty} p_x$ and substituting $y = x - 1$, we note that

$$T = p + \sum_{x=2}^{\infty} q\,p_{x-1} = p + q\,\sum_{y=1}^{\infty} p_y = p + q\,T.$$

Solve this relation to get $T = p/(1-q) = 1$.

(G2) PGF: Writing $P_X(t) = \sum_{x=1}^{\infty} p_x t^x$ and substituting $y = x - 1$, we have

$$P_X(t) = pt + \sum_{x=2}^{\infty} q\,p_{x-1} t^x = pt + qt\,\sum_{y=1}^{\infty} p_y t^y = pt + qtP_X(t),$$

which we solve to get $P_X(t) = pt\,(1-qt)^{-1}$, for $0 < t < q^{-1}$.

(G3) First moment (or mean): Taking expectations on the two sides of (2), we have

$$E[X] = p + q\,(1 + E[\tilde{X}]) = 1 + q\,E[X],$$

solving which we get $E[X] = p^{-1}$.

(G4) MGF: Writing $M_X(t) = \sum_{x=1}^{\infty} p_x\,e^{tx}$ and substituting $y = x - 1$, we have

$$M_X(t) = p\,e^t + \sum_{x=2}^{\infty} q\,p_{x-1} e^{tx} = p\,e^t + qe^t\,\sum_{y=1}^{\infty} p_y e^{ty} = p\,e^t + q\,e^t\,M_X(t).$$

Hence, we get $M_X(t) = pe^t\,(1-qe^t)^{-1}$, for $0 < t < -\ln q$.

(G5) Lack of memory property: Every time a tail (T) occurs, the process renews itself. The event $\{X > x\}$ means the first x tosses are all T's. Likewise, the conditional event $\{X > x + y | X > x\}$ means that when the first x tosses are known to be all T's, the next y tosses also turn out to be all T's. However, the second batch of y tosses is independent of the first batch of x tosses. Hence, the desired conditional probability $P\{X > x + y | X > x\}$ equals the unconditional probability $P\{X > y\}$.

Immediate extension: The geometric(p) RV is also called a negative binomial($1, p$) RV. If we change the definition of success from "the appearance of the first head" to "the appearance of the r-th head" (for $r \geq 2$) or until we have a cumulative total of r heads, then the waiting time $Y = X_1 + \ldots + X_r$, where $X_1, X_2, \ldots, X_r$ are independent geometric(p) RVs, is said to have a negative binomial(r, p) distribution. Its first moment is r/p; its PGF is $P_Y(t) = (pt)^r\,(1-qt)^{-r}$, for $0 < t < q^{-1}$; and its MGF is $M_Y(t) = (pe^t)^r\,(1-qe^t)^{-r}$, for $0 < t < -\ln q$.

1.1. Relation to existing literature

Waiting-time problems for patterns in random sequences have a long history in probability theory. Classical approaches include martingale methods, generating functions, finite-state Markov chain embeddings, and renewal-theoretic techniques. For example, Li (1980) studied the occurrence of patterns using martingale arguments, while Ebneshahrashoob and Sobel (1990) focused on the sooner and later waiting-time problems in Bernoulli sequences. Fu and Chang (2002) and Uchida (1998) developed generating-function methods for waiting times associated with compound and multistate patterns.

The present paper has a different objective. Rather than seeking the most general theorem for a particular class of patterns, we emphasize a unified recursive framework based on one-step transition probabilities and renewal relations. The same methodology is used throughout the paper to derive exact PMFs for waiting times arising in independent trials, multinomial experiments, finite Markov chains, and random walks on graphs.

The novelty of the present paper is not in the examples considered, but in the use of a common recursive framework based on renewal relations and one-step transition probabilities. Repeated matrix recursion extends the familiar recursive construction of the geometric distribution. This perspective yields explicit PMFs for waiting times in several seemingly different settings and highlights their structural connections. The paper serves both as a study of waiting-time distributions and as an exposition of a unifying probabilistic method.

1.2. Motivation and applications

Waiting-time distributions arise naturally in a wide variety of scientific settings. In quality control, one may be interested in the number of inspections preceding a prescribed sequence of outcomes. In reliability and communications, the occurrence of particular patterns of successes and failures may signal the onset of an important event. In bioinformatics, researchers study waiting times until specified nucleotide patterns appear in DNA sequences. In stochastic processes, return times in Markov chains and random walks are fundamental quantities related to recurrence, mixing behavior, and long-run structure. While expected waiting times are often reported, the full PMF provides substantially richer information by quantifying variability and tail behavior. The examples studied in this paper illustrate how a common recursive methodology can be used to derive exact waiting-time distributions in several such settings. In statistical applications, waiting-time distributions provide benchmark models against which simulation studies and empirical observations can be compared.

The rest of the paper considers other, more complex notions of success. The paper is organized as follows: In Sections 2–4, we study the waiting time until a specified block of outcomes appears in (i) a sequence of independent, binary tosses, (ii) in a sequence of independent, multinomial outcomes, and (iii) in a Markov chain. In Section 5, we study the time until a random walk on the vertices of a graph returns to its starting position. Section 6 concludes the paper with a summary and several open problems.

2. Until a specific block of binary outcomes occurs

Consider a sequence of independent, binary RVs in which the appearance of a specific block of outcomes (for example, HTH or HTT or HTHTH) signifies a success. A popular

brain teaser goes as follows: "Consider tossing a *fair* coin repeatedly. Let a denote the expected number of tosses until the pattern HTH occurs, and b until HTT occurs. Which of the three cases holds true: $a < b$, $a = b$, or $a > b$?"

Many readers initially conjecture that $a = b$. Refer to Donnelly (2005). But actually, $a > b$. Here is a layman's justification: When you are waiting for either pattern, you are somewhat excited to see the first H; then you are even more excited to see the pattern HT emerge. Now, if you were waiting for the pattern HTH, but instead HTT occurs, all hope is lost, and the process begins afresh for you. However, if you were waiting for HTT, but HTH occurs, not all is lost! You still have some excitement left because you already see the initial H of the desirable pattern. In fact, $a = 10, b = 8$; the difference of 2 is exactly the expected waiting time until the first H occurs. In Lemma 1 below, we give a rigorous proof not just for the fair coin, but for the more general case of a possibly biased coin.

It is worth repeating the key idea: Partial progress toward a target pattern may or may not be retained after a mismatch occurs. For HTT, the appearance of HTH still preserves the initial H needed for a future occurrence of HTT. By contrast, when searching for HTH, the appearance of HTT destroys all progress toward the target pattern. The recursive equations below quantify this phenomenon.

Lemma 1: The expected number of tosses of a biased coin with $P(\text{Heads}) = p$ until the pattern HTH occurs is $a = p^{-2}q^{-1} + p^{-1}$, and that until HTT occurs is $b = p^{-1}q^{-2}$.

Figure 1 shows the schematic diagram of the patterns of outcomes until success.

Figure 1: Transitions until a specific block of outcomes occurs: (a) HTH, and (b) HTT (each arrow represents 1/2 probability)

Proof of Lemma 1 For any (partial) pattern Z such as ϕ, H, T, HT, etc., let $\mu(Z)$ denote the expected number of additional tosses until the target pattern HTH appears, starting after the appearance of the pattern Z. Clearly, $\mu(\phi) = a$, and the following system of equalities hold:

$$\begin{pmatrix} a \\ \mu(H) \\ \mu(HT) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{bmatrix} q & p & 0 \\ 0 & p & q \\ q & 0 & 0 \end{bmatrix} \begin{pmatrix} a \\ \mu(H) \\ \mu(HT) \end{pmatrix},$$

solving which we get

$$\begin{pmatrix} a \\ \mu(H) \\ \mu(HT) \end{pmatrix} = \begin{bmatrix} p & -p & 0 \\ 0 & q & -q \\ -q & 0 & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{p^2 q} \begin{pmatrix} 1+pq \\ 1 \\ (1+p)q \end{pmatrix}. \tag{4}$$

In particular, $a = (1 + pq)/(p^2q) = p^{-2}q^{-1} + p^{-1}$. Incidentally, from the second and third coordinates in (4), we get $\mu(H) = p^{-2}q^{-1}$ and $\mu(HT) = p^{-2} + p^{-1}$.

Likewise, for any pattern Z, let $\nu(Z)$ denote the expected number of additional tosses until the target pattern HTT appears, after the appearance of Z. Then $\nu(\phi) = b$, and

$$\begin{pmatrix} b \\ \nu(H) \\ \nu(HT) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{bmatrix} q & p & 0 \\ 0 & p & q \\ 0 & p & 0 \end{bmatrix} \begin{pmatrix} b \\ \nu(H) \\ \nu(HT) \end{pmatrix},$$

This system of equations decomposes into: $b = 1 + p\ \nu(H) + q\ b = p^{-1} + \nu(H)$ and

$$\begin{pmatrix} \nu(H) \\ \nu(HT) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{bmatrix} p & q \\ p & 0 \end{bmatrix} \begin{pmatrix} \nu(H) \\ \nu(HT) \end{pmatrix},$$

which, when solved, gives us

$$\begin{pmatrix} \nu(H) \\ \nu(HT) \end{pmatrix} = \begin{bmatrix} q & -q \\ -p & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{q^2} \begin{pmatrix} 1+q \\ 1 \end{pmatrix}. \tag{5}$$

That is, $\nu(H) = q^{-2} + q^{-1}$ and $\nu(HT) = q^{-2}$. Thereafter, we obtain $b = p^{-1} + \nu(H) = p^{-1} + q^{-2} + q^{-1} = (pq)^{-1} + q^{-2} = p^{-1}q^{-2}$. The proof is now complete. Q.E.D.

Lemma 1 can be generalized to the following proposition. See Ross (1996).

Theorem 1: Suppose that the pattern Z, consisting of i H's and j T's, when the **fewest** number of leading symbols is deleted, becomes a sub-pattern $\tilde{Z}$ that is the same as when that many trailing symbols are deleted from Z. Then the expected number of tosses until Z appears is $p^{-i}q^{-j}$ more than the expected number of tosses until $\tilde{Z}$ appears.

For example, the expected time until pattern HTHTH appears is

$$\mu(HTHTH) = p^{-3}q^{-2} + \mu(HTH) = p^{-3}q^{-2} + p^{-2}q^{-1} + \mu(H) = p^{-3}q^{-2} + p^{-2}q^{-1} + p^{-1}.$$

For more general results regarding: (1) the expected time until a pattern occurs starting from another given pattern, (2) the expected time until one of several patterns occurs, and (3) the probability that one pattern occurs before some other patterns, see Li (1980).

2.1. Waiting time until HTT

So far, in this section, we studied the *expected* waiting time until a pattern appears. What about the PMF of the waiting time? Indeed, we can derive the PMF by suitably utilizing the renewal property. That is, having studied a property analogous to (G3), let us now focus on a property analogous to (G1).

Let X denote the waiting time until the pattern HTT occurs. Also, let $X(H)$ and $X(HT)$ denote the number of additional tosses until the target pattern HTT appears, starting after the appearance of H and HT, respectively. Then in the spirit of (2), we have

$$X = \begin{cases} 1 + X(H) & \text{with probability } p \\ 1 + \tilde{X} & \text{with probability } q, \end{cases} \tag{6}$$

where $\tilde{X}$ has the same distribution as X. Also, $X(H)$ and $X(HT)$ are interrelated as

$$X(H) = \begin{cases} 1 + \tilde{X}(H) & \text{w.p. } p \\ 1 + X(HT) & \text{w.p. } q; \end{cases} \quad \text{and} \quad X(HT) = \begin{cases} 1 + X(H) & \text{w.p. } p \\ 1 & \text{w.p. } q; \end{cases} \tag{7}$$

where $\tilde{X}(H)$ has the same distribution as $X(H)$. The two equations in (7) imply that

$$X(H) = \begin{cases} 1 + \tilde{X}(H) & \text{w.p. } p \\ 2 + \tilde{X}(H) & \text{w.p. } qp \\ 2 & \text{w.p. } q^2. \end{cases} \tag{8}$$

Thus, the distribution of $X(H)$ is related to itself, or the renewal property holds. Let the PMF of $X(H)$ be $\mathbf{p}_{X(H)} = (p_1, p_2, p_3, \ldots)$. From (8), we have $p_1 = 0, p_2 = q^2$, and

$$p_{n+2} = p \cdot p_{n+1} + qp \cdot p_n, \text{ for } n \geq 1.$$

Let us write $\begin{pmatrix} p_2 \\ p_1 \end{pmatrix} = \begin{pmatrix} q^2 \\ 0 \end{pmatrix}$, and thereafter, for $k \geq 2$, we have

$$\begin{pmatrix} p_{k+1} \\ p_k \end{pmatrix} = \begin{bmatrix} p & pq \\ 1 & 0 \end{bmatrix} \begin{pmatrix} p_k \\ p_{k-1} \end{pmatrix} = \ldots = \begin{bmatrix} p & pq \\ 1 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} q^2 \\ 0 \end{pmatrix}.$$

Thus, we obtain the PMF $\mathbf{p}_{X(H)}$ of RV $X(H)$ by repeated pre-multiplication by a matrix, similar to how in (1) the geometric(p) probabilities were obtained by repeated pre-multiplication by q. We can verify that the total probability is

$$\sum_{k=1}^{\infty} \begin{pmatrix} p_{k+1} \\ p_k \end{pmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} p & pq \\ 1 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} q^2 \\ 0 \end{pmatrix} = \begin{bmatrix} q & -pq \\ -1 & 1 \end{bmatrix}^{-1} \begin{pmatrix} q^2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Thereafter, from the second equation within (7), we can obtain the PMF of $X(HT)$ as $(q, p \cdot \mathbf{p}_{X(H)})$; and from (6), the PMF of X is $\mathbf{p}_X = (0, p \cdot \mathbf{p}_{X(H)} + q \cdot \mathbf{p}_X)$, or

$$pq^2\,(0, 0, 1, 1, 1, 1 - pq^2, 1 - 2pq^2, 1 - 3pq^2, \ \ldots).$$

2.2. Waiting time until HTH

Let Y denote the waiting time until the pattern HTH occurs; and let $Y(H)$ and $Y(HT)$ denote the number of additional tosses until HTH appears, after H and HT, respectively. Then writing $\mathbf{q}_k = (P\{Y(H) = k\}, P\{Y = k\}, P\{Y = k - 1\})^T$, a vector of three probability masses, we have $\mathbf{q}_2 = (pq, 0, 0)^T$, and for $k \geq 2$, we have

$$\mathbf{q}_{k+1} = \begin{bmatrix} p & 0 & q^2 \\ p & q & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{q}_k = \ldots = \begin{bmatrix} p & 0 & q^2 \\ p & q & 0 \\ 0 & 1 & 0 \end{bmatrix}^{k-1} \mathbf{q}_2.$$

Again, note the similarity between $\mathbf{q}_{k+1}$ and the mass of a geometric(p) RV in (1)! Clearly, $P\{Y = 1\} = 0$, and for $k \geq 2$, $P\{Y = k\}$ is the second element of $\mathbf{q}_k$. Hence, the entire PMF of Y, written as a vector, is

$$p^2 q\,(0, 0, 1, 1, p^2 + q, p^3 + q, p^4 + q, p^4 + [1 + p + 2p^2]q^2, \ldots).$$

Again, we can verify that the total probability is

$$\sum_{k=2}^{\infty} \mathbf{q}_k = \sum_{k=2}^{\infty} \begin{bmatrix} p & 0 & q^2 \\ p & q & 0 \\ 0 & 1 & 0 \end{bmatrix}^{k-2} \begin{pmatrix} pq \\ 0 \\ 0 \end{pmatrix} = \begin{bmatrix} q & 0 & -q^2 \\ -p & p & 0 \\ 0 & -1 & 1 \end{bmatrix}^{-1} \begin{pmatrix} pq \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

An extension: If success is redefined as the occurrence of either HTH or HTT, then the waiting time until success is exactly *one more than* the waiting time until HT occurs. As such, the waiting time to success has a PMF given by

$$\begin{aligned} &pq\,(0, 0, 1, p+q, p^2+q, p^3+q(p^2+q), p^4+q(p^2+q^2), p^5+q(p^4+p^2q+q^3), \ldots) \\ &= pq\,(0, 0, 1, 1, (p^2+q), (p^2+q^2), (p^4+p^2q+q^3), (p^4+p^2q^2+q^4), \ldots). \end{aligned}$$

We leave it to the reader to check that the sum of the probabilities is one.

3. Until a block of multinomial outcomes occurs

Each coin toss results in a binary outcome. What if the outcome were multinomial?

Several authors studied the waiting-time distributions for multinomial patterns. Uchida (1998) employed generalized PGFs to study waiting times and counts of pattern occurrences. Fu and Chang (2002) used finite Markov chain embedding techniques to derive PGFs and develop computational procedures for compound patterns. In contrast, the present paper emphasizes direct recursive construction of the probability mass function through one-step transition probabilities. This approach naturally leads to matrix recursions analogous to the recursion defining the geometric distribution and provides a common framework applicable throughout the paper.

As an example, consider the popular game of "Rock-Paper-Scissors." The main objective of this section is not the game itself, but to illustrate how the recursive framework extends naturally from binary outcomes to multinomial outcomes. The game provides a familiar setting in which several competing patterns may define success.

In Rock-Paper-Scissors, two players simultaneously display with their hands one of these three symbols R-P-S. If both players display the same symbol, the play ends in a tie, and no one scores any points; otherwise, rock loses to paper, paper loses to scissors, and scissors loses to rock, and the winner scores one point. What are the probabilities of the three outcomes—a tie, A wins, and B wins? The probabilities depend on the strategies the players employ. We will consider several models of their strategies.

Scenario 1: Suppose that Player A selects the symbols R-P-S independently at random with probabilities $(.3, .4, .3)$, Player B selects the symbol independently at random with probabilities $(.2, .4, .4)$; and neither player knows the probability distribution of the other. Below, we tabulate the joint probabilities of the two players' selections, with the rows representing Player A's choices, the columns Player B's, and the superscripts telling us who wins

(here T denotes a tie):

$$\begin{array}{ccc|cc} .06^T & .12^B & .12^A & .3 & R \\ .08^A & .16^T & .16^B & .4 & P \quad \text{A's choice} \\ .06^B & .12^A & .12^T & .3 & S \\ \hline .2 & .4 & .4 & & \\ R & P & S & & \\ \multicolumn{3}{c}{\text{B's choice}} & & \end{array}$$

In this case, the play ends in a tie with probability $.06 + .16 + .12 = .34$, A wins the play with probability $.12 + .08 + .12 = .32$, and B wins with probability $.12 + .16 + .06 = .34$. Thus, the game is slightly advantageous to Player B.

Scenario 2: Suppose Player A selects the symbols R-P-S independently at random with probabilities $(.3, .4, .3)$, and Player B knows this strategy. To beat Player A's most likely choice P, Player B selects the symbol S all the time. In this case, the play ends in a tie with probability .3, A wins with probability .3 and B wins with probability .4. Thus, by utilizing her knowledge of A's strategy, B substantially increases her advantage over A.

However, B's advantage is not sustainable in the long run. As soon as Player A would recognize Player B's strategy (of always choosing S), he ought to switch his own strategy to always choosing R; and thereby always winning.

Scenario 3: Suppose that Player A selects the symbols R-P-S independently at random with probabilities $(1/3, 1/3, 1/3)$, and declares it to Player B. Then, no matter which probability distribution Player B uses, the play ends in a tie, or A wins, or B wins with probability 1/3 each. Not to give Player A any potential advantage by discovering her strategy, Player B should choose R-P-S with equal probabilities also.

In the rest of this subsection, we assume that Scenario 3 holds; that is, we suppose that at least one player chooses R-P-S with equal probability in each play. We address three questions: How many plays are needed until

(Q1) Player A wins three in a row (or the pattern AAA appears)?

(Q2) either Player A or Player B wins three in a row (either AAA or BBB appears)? and

(Q3) the same outcome occurs three times in a row (either AAA or BBB or TTT occurs)?

Figure 2 shows the recursive relations involved in answering the three questions.

Answer to (Q1):

Let N_{A^3} denote the number of plays until Player A wins three in a row. Using Proposition 2.1, $E[N_{A^3}] = 3^3 + 3^2 + 3 = 39$. To find the PMF of N_{A^3}, let us write

$$\mathbf{q}_k = (P\{N_{A^3} = k\}, P\{N_{A^3}(A) = k\}, P\{N_{A^3}(AA) = k\})^T.$$

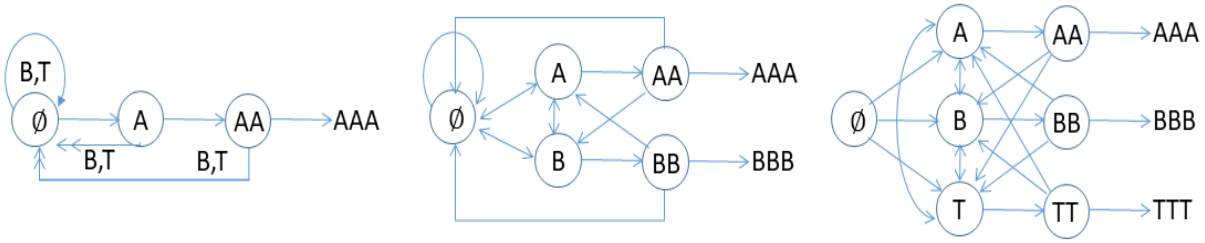


Figure 2: How many plays until (a) A^3, (b) $\{A^3, B^3\}$, and (c) $\{A^3, B^3, T^3\}$ occurs? (Each arrowhead represents 1/3 probability of transition in that direction.)

Then, we have

$$\mathbf{q}_1 = \begin{pmatrix} 0 \\ 0 \\ 1/3 \end{pmatrix} \text{ and for } k \geq 1, \ \mathbf{q}_{k+1} = \begin{bmatrix} 2/3 & 1/3 & 0 \\ 2/3 & 0 & 1/3 \\ 2/3 & 0 & 0 \end{bmatrix} \mathbf{q}_k,$$

solving which, we have

$$p_k = P\{N_{A^3} = k\} \ = \ 3^{-k} \ (1,0,0) \begin{bmatrix} 2 & 1 & 0 \\ 2 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \text{ for } k \geq 1.$$

Note the similarity to the PMF of the geometric(p) distribution in (1). In particular, N_{A^3} has PMF $(0, 0, 1/3^3, 2/3^4, 6/3^5, 18/3^6, 52/3^7 \ldots)^T$; or,

$$p_k = P\{N_{A^3} = k\} = 3^{-k} a_k,$$

where $a_1 = 0, a_2 = 0, a_3 = 1$ and $a_{k+1} = 2(a_k + a_{k-1} + a_{k-2})$ for $k \geq 3$.

Answer to (Q2):

Let N_{A^3,B^3} denote the number of plays until either player wins three in a row. Then $E[N_{A^3,B^3}] = 39/2$. To find the PMF of N_{A^3,B^3}, note that $N_{A^3,B^3}(A)$ and $N_{A^3,B^3}(B)$ have identical distributions, and so do $N_{A^3,B^3}(AA)$ and $N_{A^3,B^3}(BB)$. Writing

$$\mathbf{q}_k = (P\{N_{A^3,B^3} = k\}, P\{N_{A^3,B^3}(A) = k\}, P\{N_{A^3,B^3}(AA) = k\})^T,$$

we have

$$\mathbf{q}_1 = \begin{pmatrix} 0 \\ 0 \\ 1/3 \end{pmatrix} \text{ and for } k \geq 1, \ \mathbf{q}_{k+1} = \begin{bmatrix} 1/3 & 2/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 \end{bmatrix} \mathbf{q}_k.$$

Therefore, the PMF of N_{A^3,B^3} is given by

$$p_k = P\{N_{A^3,B^3} = k\} \ = \ 3^{-k} \ (1,0,0) \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \text{ for } k \geq 1.$$

Again, note the similarity to the PMF of the geometric(p) distribution in (1). In particular, N_{A^3,B^3} has PMF $(0, 0, 2/3^3, 4/3^4, 12/3^5, 34/3^6, 96/3^7 \ldots)^T$; or,

$$p_k = P\{N_{A^3,B^3} = k\} = 3^{-k} b_k,$$

where $b_1 = 0, b_2 = 0, b_3 = 1$, and $b_{k+1} = 2(b_k + b_{k-1}) + b_{k-2}$ for $k \geq 3$.

Answer to (Q3):

Let N denote the number of plays until one of the patterns AAA or BBB or TTT appears. Then $E[N] = 39/3 = 13$. To find the PMF $\mathbf{p}$ of N, note that $N(A), N(B), N(T)$ have identical distributions, and so do $N(AA), N(BB), N(TT)$. Writing

$$\mathbf{q}_k = (P\{N(A) = k\}, P\{N(AA) = k\})^T,$$

we have

$$\mathbf{q}_1 = \begin{pmatrix} 0 \\ 1/3 \end{pmatrix} \text{ and for } k \geq 1, \ \mathbf{q}_{k+1} = \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 0 \end{bmatrix} \mathbf{q}_k.$$

Therefore, the PMF $\mathbf{p}$ of N is given by $p_1 = 0$, and

$$p_{k+1} = P\{N = k+1\} = P\{N(A) = k\} = 3^{-k} \ (1, 0) \begin{bmatrix} 2 & 1 \\ 2 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \text{ for } k \geq 1.$$

Again, note the similarity to the PMF of the geometric(p) distribution in (1). Specifically, the PMF of N is $\mathbf{p} = (0, 0, 1/3^2, 2/3^3, 6/3^4, 16/3^5, 44/3^6, \ldots)$; or,

$$p_k = P\{N = k\} = 3^{-(k-1)} c_k,$$

where $c_1 = 0, c_2 = 0, c_3 = 1$, and thereafter $c_{k+1} = 2(c_k + c_{k-1})$ for $k \geq 3$.

Another extension: Suppose that the game ends as soon as one player earns 10 points, and that player receives as many dollars as her points exceed the other player's. What is the PMF of the number of plays N_{10} when the game ends? What is the PMF of the winnings W_{10}^A of Player A? Who will win the game?

The number of plays until Player A earns 10 points has a negative binomial$(10, 1/3)$ distribution, which is the sum of 10 independent, geometric$(1/3)$ RVs. However, the distribution of N_{10}, the time until either Player A or Player B earns 10 points (the new definition of success), is not commonly known. Let us call its distribution the *doubly negative binomial*$(10, 10; 1/3, 1/3)$. The attentive reader should check that:

(i) The PMF of N_{10} is $P\{N_{10} = 10 + k + t\} = 2\,(9 + k + t)!/[9!k!t!]\, 3^{-(10+k+t)}$, where k denotes the points earned by the loser (with $0 \leq k \leq 9$) and $t \geq 0$ denotes the number of ties during the game.

(ii) The PMF of W_{10}^A is symmetric over $\{\pm 1, \pm 2, \ldots, \pm 10\}$; and it is given by

$$P\{W_{10}^A = -j\} = P\{W_{10}^A = j\} = \binom{19 - j}{9} 2^{-(20-j)} \text{ for } j = 1, 2 \ldots, 10.$$

(iii) By symmetry, it follows that $P\{A \text{ wins}\} = 1/2 = P\{B \text{ wins}\}$.

4. Until a Markov chain exhibits a desired feature

So far we have considered sequences of outcomes that are independent. What if they are not? As the simplest departure from independence, we consider Markov chains. In a Markov chain, the probability distribution of the next outcome depends only on the current outcome and is independent of all previous outcomes.

For example, in Rock-Paper-Scissors, suppose that Player A chooses R-P-S according to a probability distribution that depends on his or her latest choice, but not on any previous choices. Suppose that the conditional probabilities of the next choice, given the current choice, are as follows: If the latest choice was R, then the next choice will be R-P-S with respective probabilities (.1, .5, .4). If the latest choice was P, then the probabilities are (.4, .3, .3). Finally, if the latest choice was S, then the probabilities are (.3, .2, .5). We summarize all such information in a one-step transition probability matrix

$$P = \begin{array}{c} \\ \begin{pmatrix} .1 & .5 & .4 \\ .4 & .3 & .3 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{matrix} R & P & S \end{matrix} \end{array} \quad \begin{array}{l} R \\ P \\ S \end{array} \text{ A's current choice}$$

A's next choice

where P_{ij} denotes the conditional probability that A's next choice is j, given that her current choice is i, and is free of all previous choices.

The above example is a random walk on three states {R, P, S}={rock, paper, scissors}. (The graph includes loops at each vertex). The given information suffices to determine what proportion of choices are (R, P, S) in the long run. Here is how: Reasoning backwards, we say that if such a probability distribution π exists and if the latest choice was made with probability distribution π, then the next choice will also have the same distribution π; that is, $\pi^T P = \pi^T, \pi^T \mathbf{1} = 1$, where $\mathbf{1}$ denotes the vector of all entries 1. Therefore, it suffices to find the eigen vector of the matrix $(I - P^T)$ corresponding to the eigenvalue 0, and standardize it to ensure the sum is one. Alternatively, we may compute P^n for a sufficiently large $n \geq 1$ (in fact, to be more efficient, we compute $P^2, P^4, P^8, \ldots$) until its rows are (almost) identical. For the above P matrix, the answer is $\pi_A^T = (.2762, .3143, .4095)$.

Suppose that Player B chooses R-P-S according to another Markov chain with transition matrix

$$Q = \begin{array}{c} \\ \begin{pmatrix} .3 & .4 & .3 \\ .2 & .3 & .5 \\ .4 & .4 & .2 \end{pmatrix} \\ \begin{matrix} R & P & S \end{matrix} \end{array} \quad \begin{array}{l} R \\ P \\ S \end{array} \text{ B's current choice}$$

B's next choice

Then B chooses R-P-S with marginal stationary probabilities $\pi_B^T = (.2975, .3636, .3388)$.

If each of Players A and B acts according to his or her own transition matrix, but independently of each other's choice, then in the long run the probabilities of a tie, of A winning, and of B winning can be obtained as was done in Scenario 1, with marginal selection probabilities replaced by their respective stationary probabilities. The reader may

verify that B's probabilities of win, loss, and tie are .3288, .3360, and .3352, respectively. Thus, the game is slightly unfavorable to Player B.

However, suppose that after playing for a long time, Player B analyzed Player A's behavior and discovered his transition matrix P. (To do so, simply use the relative frequencies of RR-RP-RS to estimate the first row of P, those of PR-PP-PS to estimate the second row, and those of SR-SP-SS to estimate the third row.) However, Player A remained oblivious to Player B's transition matrix Q. How should Player B take advantage of her knowledge?

Since whatever option A chooses in one play determines his probability distribution in the next play, B should choose that option which will beat A's most likely next choice. In other words, if A chose R/P/S in the current play, then in the next play B should choose respectively S/P/R. With this strategy, in the next play B will win with respective probabilities $(.5, .4, .5)$, and lose with respective probabilities $(.1, .3, .2)$ and the game will end up in a tie with respective probabilities $(.4, .4, .3)$. Overall, B's probabilities of win, loss and tie are .4686, .2038 and .3276 respectively. The game became highly advantageous to Player B who analyzed Player A's past behavior to discover his transition matrix. It pays to analyze the opponent's strategy!

5. Until a random walk returns to origin

The examples in this section demonstrate that the same recursive methodology applies beyond pattern-occurrence problems. Return times in random walks are traditionally studied through expectations and asymptotic methods; here we construct their full distributions by identifying suitable renewal structures.

In general, a random walk is a discrete-time Markov chain; that is, a sequence of states in which the next state is chosen according to a conditional probability distribution that depends on the current state. For a symmetric random walk on the vertices of a graph, the next vertex is equally likely to be any one of the adjacent vertices. Random walks on graphs are related to reversible Markov chains (see Aldous and Fill (2014)) and to electric networks (see Doyle and Snell (1984)). We focus on symmetric random walks on a linear path, a cycle, a wheel-and-spokes graph, and some familiar solids such as a tetrahedron, an octahedron, and a hexahedron.

At time 0, the walk is at a particular vertex, which we call the starting vertex or the origin. We will study the time until the walk returns to the origin. Much of the literature on random walks (see Sarkar (2006); Maiti and Sarkar (2019)) emphasizes expectations, variances, recurrence properties, and asymptotic behavior of return times. Motivated by these studies, we examine the full PMF of the return time for several finite graphs and show how it can be obtained recursively from one-step transition probabilities.

5.1. Random walks on a linear path, a cycle, or a wheel-and-spokes

Here we study a symmetric random walk on some simple graphs. See Figure 3.

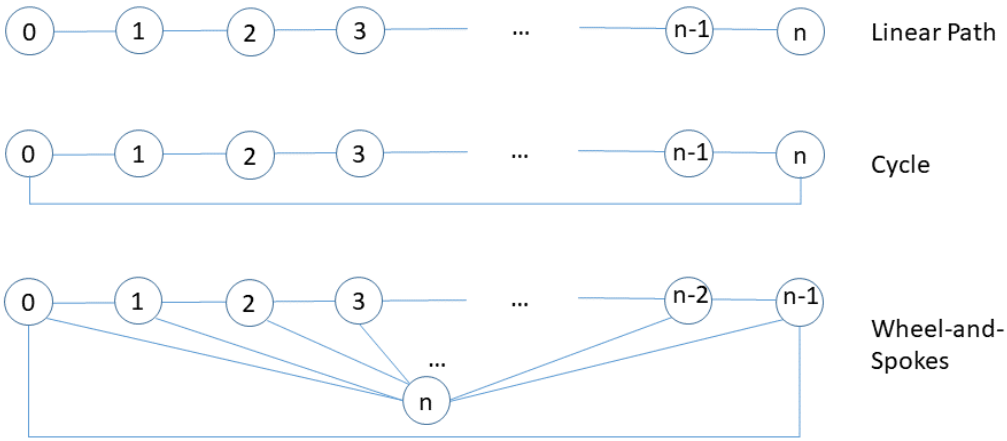


Figure 3: Graphs of (a) a linear path L_n, (b) a cycle C_n, and (c) a wheel-and-spokes W_n. Transition on each edge happens in both directions.

5.1.1. *Linear path*

A linear path L_n is a graph consisting of vertices $\{0, 1, 2, \ldots, n\}$ and edges $(i-1, i)$ for $i = 1, 2, \ldots, n$. The walk begins at vertex 0 at time $t = 0$. At time $t = 1$, the walk surely reaches vertex 1. Whenever the walk is at a vertex i (for $i = 1, 2, \ldots, n-1$), at the next instance, it moves to either its left neighbor $(i-1)$ or its right neighbor $(i+1)$, with probability $1/2$ each. Finally, from vertex n the walk surely moves to vertex $(n-1)$. Let T_R denote the waiting time until the walk returns to the starting vertex 0. It is well known that $E[T_R] = n$. Here, we derive the PMF of T_R.

Let ${}_iT_j$ denote the number of steps needed for the walk to go from vertex i to vertex j. Then $T_R = 1 + {}_1T_0$; and the following recursive relations hold:

$$
{}_1T_0 = \begin{cases} 1 & \text{w.p. } 1/2 \\ 1 + {}_2T_0 & \text{w.p. } 1/2 \end{cases}, \quad {}_2T_0 = \begin{cases} 1 + {}_1T_0 & \text{w.p. } 1/2 \\ 1 + {}_3T_0 & \text{w.p. } 1/2 \end{cases}, \quad \ldots,
$$
$$
{}_{n-1}T_0 = \begin{cases} 1 + {}_{n-2}T_0 & \text{w.p. } 1/2 \\ 1 + {}_nT_0 & \text{w.p. } 1/2 \end{cases},
$$

and ${}_nT_0 = 1 + {}_{n-1}T_0$. Thus, to find the PMF of T_R, we need the PMF of ${}_1T_0$, for which we need the PMF of ${}_2T_0$, etc. In fact, we can simultaneously find the PMF of the entire random vector $({}_1T_0,\ {}_2T_0,\ {}_3T_0, \ldots,\ {}_nT_0)$. To do so, define the following $n \times 1$ vectors

$$
\mathbf{p}_k = (P\{\ {}_1T_0 = k\},\ P\{\ {}_2T_0 = k\},\ P\{\ {}_3T_0 = k\}, \ldots,\ P\{\ {}_nT_0 = k\})^T, \text{ for } k \geq 1.
$$

Then $\mathbf{p}_1 = 2^{-1}(1, 0, \ldots, 0)^T$, and for $k \geq 2$,

$$
\mathbf{p}_k = R\,\mathbf{p}_{k-1} = \ldots = R^{k-1}\,\mathbf{p}_1, \tag{9}
$$

where R is an $n \times n$ matrix given by

$$R = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & \dots & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & \dots & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ \dots & & \dots & & & \dots & \\ 0 & & \dots & 0 & 1/2 & 0 & 1/2 \\ 0 & & \dots & 0 & 0 & 1 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 1 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ \dots & & \dots & & & \dots & \\ 0 & & \dots & 0 & 1 & 0 & 1 \\ 0 & & \dots & 0 & 0 & 2 & 0 \end{bmatrix}.$$

Note that $\mathbf{p}_k$ in (9) is formally similar to a geometric(p) RV given in (1). Using (9), we evaluate $\mathbf{p}_2 = 2^{-2}(0, 1, 0, 0, \dots, 0)^T$, $\mathbf{p}_3 = 2^{-3}(1, 0, 1, 0, \dots, 0)^T$, etc. The attentive reader should verify that $\sum_{k=1}^{\infty} \mathbf{p}_k = (I - R)^{-1}\mathbf{p}_1 = \mathbf{1}$, the $n \times 1$ vector of all entries 1.

5.1.2. *Cycle*

To a linear path L_n, let us augment an edge joining vertices n and 0. See Figure 3 (middle panel). The resulting graph C_n is called the cycle graph consisting of vertices $\{0, 1, 2, \dots, n\}$ on a circle numbered clockwise. Suppose that at time $t = 0$, the walk begins at vertex 0. Whenever the walk is at a vertex i (for $i = 0, 1, 2, \dots, n$), at the next instance it moves to either its counterclockwise neighbor $(i - 1)$ or its clockwise neighbor $(i + 1)$ with probability 1/2 each. Here, addition and subtraction are done modulo $(n + 1)$. We are interested in the waiting time T_R until the walk returns to the starting vertex 0. Again, it is well known that $E[T_R] = n + 1$. Here, we derive the full PMF of T_R.

Without loss of generality, we assume that the walk goes to vertex 1 at time $t = 1$ (for otherwise, we can reverse-number the vertices). Also, splitting vertex 0 into two vertices 0 and $(n + 1)$, we can convert cyclic graph C_n into linear path L_{n+1}. Let us write $\mathbf{0}$ to denote the subset of vertices $\{0, n + 1\}$, and ${}_iT_{\mathbf{0}}$ to denote the number of steps needed for the walk to go from vertex i to either vertex 0 or vertex $n + 1$ on the linear path L_{n+1}. Then T_R on the cycle graph C_n is the same as $1 + {}_1T_{\mathbf{0}}$ on linear path L_{n+1}; and the following recursive relations hold:

$${}_1T_{\mathbf{0}} = \begin{cases} 1 & \text{w.p. } 1/2 \\ 1 + {}_2T_{\mathbf{0}} & \text{w.p. } 1/2 \end{cases}, \quad {}_2T_{\mathbf{0}} = \begin{cases} 1 + {}_1T_{\mathbf{0}} & \text{w.p. } 1/2 \\ 1 + {}_3T_{\mathbf{0}} & \text{w.p. } 1/2 \end{cases}, \quad \dots,$$

$${}_nT_{\mathbf{0}} = \begin{cases} 1 + {}_{n-1}T_{\mathbf{0}} & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2. \end{cases}$$

Furthermore, by reverse numbering, we note that ${}_1T_{\mathbf{0}}$ and ${}_nT_{\mathbf{0}}$ have the same distribution; so do ${}_2T_{\mathbf{0}}$ and ${}_{n-1}T_{\mathbf{0}}$; etc. Thus, if n is odd, then T_R on C_n has the same distribution as T_R on $L_{(n+1)/2}$. On the other hand, if n is even, then T_R on C_n has the same distribution as T_R on a modified linear path $L_{\lfloor n/2 \rfloor}$ that has an extra loop at vertex $\lfloor n/2 \rfloor$. In both cases, (9) holds, at least formally. However, for even n, the R matrix changes to an $\lfloor n/2 \rfloor \times \lfloor n/2 \rfloor$

matrix given by

$$R_1 = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & \dots & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & \dots & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ \dots & & \dots & & & \dots & \\ 0 & & \dots & 0 & 1/2 & 0 & 1/2 \\ 0 & & \dots & 0 & 0 & 1/2 & 1/2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 1 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ \dots & & \dots & & & \dots & \\ 0 & & \dots & 0 & 1 & 0 & 1 \\ 0 & & \dots & 0 & 0 & 1 & 1 \end{bmatrix}.$$

For this R_1 matrix also, the reader can verify that $\sum_{k=1}^{\infty} \mathbf{p}_k = (I - R_1)^{-1}\mathbf{p}_1 = \mathbf{1}$.

5.1.3. *Wheel-and-spokes*

To a cycle C_{n-1}, let us augment a new vertex n (at the center of the cycle) and join it to all vertices 0 through $(n-1)$. See Figure 3 (lower panel). The resultant graph W_n is called the wheel-and-spokes graph consisting of vertices $\{0, 1, 2, \dots, n-1\}$ on a circle and numbered clockwise, and vertex n at the center. Whenever the walk is at a vertex i (for $i = 0, 1, 2, \dots, n-1$), at the next instance it moves to either its counterclockwise neighbor $(i-1)$ or its clockwise neighbor $(i+1)$, or to the central hub with probability 1/3 each. [Again, addition and subtraction are done modulo $(n+1)$.] We are interested in the waiting time T_R until the walk returns to the starting vertex.

If the starting vertex is n, without loss of generality, we assume that the walk goes to vertex 0 at time $t = 1$ (for otherwise, we can renumber the vertices). Thereafter, the walk returns to vertex n with probability 1/3 or goes to another vertex, which we can renumber as vertex 0. Hence, T_R equals one more than a geometric(1/3) RV.

On the other hand, if the starting vertex is on the circle, we renumber it as vertex 0. Then we have the following interrelations:

$$T_R = \begin{cases} 1 + {}_nT_0 & \text{w.p. } 1/3 \\ 1 + {}_1T_0 & \text{w.p. } 2/3 \end{cases}; \quad {}_nT_0 = \begin{cases} 1 & \text{w.p. } 1/n \\ 1 + {}_iT_1 & \text{w.p. } 1/n, \text{ for } 1 \le i \le n-1 \end{cases};$$

$${}_1T_0 = \begin{cases} 1 & \text{w.p. } 1/3 \\ 1 + {}_2T_0 & \text{w.p. } 1/3 \\ 1 + {}_nT_0 & \text{w.p. } 1/3 \end{cases}; \quad {}_{n-1}T_0 = \begin{cases} 1 & \text{w.p. } 1/3 \\ 1 + {}_{n-2}T_0 & \text{w.p. } 1/3 \\ 1 + {}_nT_0 & \text{w.p. } 1/3 \end{cases};$$

$$\text{and for } 2 \le i \le n-2, \quad {}_iT_0 = \begin{cases} 1 + {}_{i-1}T_0 & \text{w.p. } 1/3 \\ 1 + {}_{i+1}T_0 & \text{w.p. } 1/3 \\ 1 + {}_nT_0 & \text{w.p. } 1/3 \end{cases}.$$

Writing ${}_ip_k = P\{ {}_iT_0 = k\}$ for $i = 1, \dots, n$, and stacking them to form a column vector $\mathbf{p}_k = ({}_1p_k, {}_2p_k, \dots, {}_np_k)^T$, we note that $\mathbf{p}_1 = (1/3, 0, 0, \dots, 0, 1/3, 1/n)^T$ and $\mathbf{p}_{k+1} =$

$R_w\,\mathbf{p}_k = \ldots = R_w^k\,\mathbf{p}_1$, formally agreeing with the geometric(p) distribution, where

$$R_w = \begin{bmatrix} 0 & 1/3 & 0 & 0 & 0 & \ldots & 0 & 1/3 \\ 1/3 & 0 & 1/3 & 0 & 0 & \ldots & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & \ldots & 0 & 1/3 \\ \ldots & & \ldots & & & \ldots & & \ldots \\ \ldots & & \ldots & & & \ldots & & \ldots \\ 0 & & \ldots & 0 & 0 & 1/3 & 0 & 1/3 \\ 1/n & & \ldots & 1/n & 1/n & 1/n & 1/n & 0 \end{bmatrix}.$$

For this R_w matrix also, the reader can verify that $\sum_{k=1}^{\infty} \mathbf{p}_k = (I - R_w)^{-1}\mathbf{p}_1 = \mathbf{1}$. Finally, $P\{T_R = k+1\} = (\ {}_1p_k + \ {}_{n-1}p_k + \ {}_1p_k)/3 = (2/3)\ {}_1p_k + (1/3)\ {}_np_k$, for $k \geq 2$.

5.2. Random walks on a tetrahedron, an octahedron, or a hexahedron

Whereas the expected return time to the origin is documented in the literature (see Sarkar (2020)), we study the return time distribution for a random walk on the vertices of three Platonic solids—tetrahedron, octahedron, and hexahedron. See Figure 4. We leave it to the reader to study the remaining two Platonic solids—dodecahedron and icosahedron. They should also ponder why there are exactly five Platonic solids. See Wi (date).

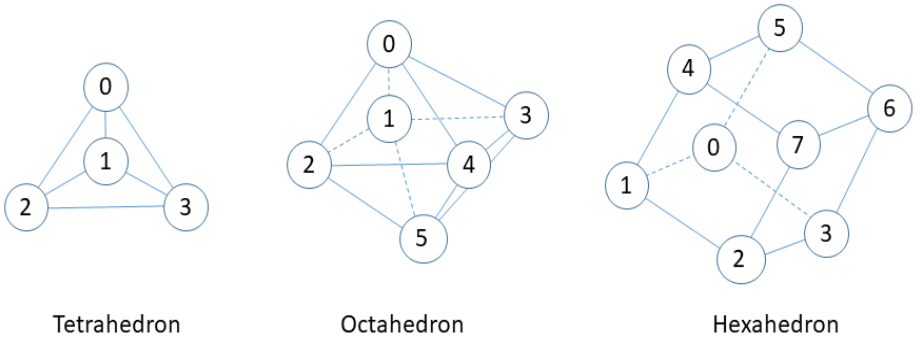


Figure 4: Graphs of (a) a tetrahedron, (b) an octahedron, and (c) a hexahedron

5.2.1. *Tetrahedron*

The graph of a tetrahedron is the complete graph on four vertices. Starting from the origin, the walk goes to any one of the other three vertices. Thereafter, at each step, the walk can return to the origin with probability 1/3, or the process renews. Hence, the waiting time until return to origin, T_R, is exactly one more than a geometric(1/3) RV.

5.2.2. *Octahedron*

The graph of an octahedron can be thought of as follows: Begin with a square whose four vertices, labeled counterclockwise as $1, 2, 4, 3$, are edge-connected to a fifth vertex, labeled 0, above the square and also to a sixth vertex, labeled 5, below the square. The graph is not complete because members within the pairs of diametrically opposite vertices, $(0, 5), (1, 4), (2, 3)$, are not directly edge-connected. At each step, the walk goes to one of its four neighbors with equal probability 1/4.

Starting from vertex 0, without loss of generality, the walk surely goes to vertex 1 (for otherwise we can rotate the octahedron suitably). Thereafter, at each step it returns to vertex 0 with probability 1/4, or it goes to vertex 5 with probability 1/4 (and in the next step it surely returns to one of vertices $1, 2, 3, 4$), or it goes to vertices 2 or 3 with probability 1/2. In these latter cases, we can renumber the vertices of the square to ensure that the walk is at vertex 1. Thus, we have $T_R = 1 + {}_1T_0$, where ${}_1T_0$ satisfies a slightly more involved renewal property than (2); namely,

$$ {}_1T_0 = \begin{cases} 1 & \text{w.p. } 1/4 \\ 1 + {}_1\bar{T}_0 & \text{w.p. } 1/2 \\ 2 + {}_1\tilde{T}_0 & \text{w.p. } 1/4 \end{cases} \tag{10} $$

where ${}_1\bar{T}_0$ and ${}_1\tilde{T}_0$ have the same distribution as ${}_1T_0$. Writing $\mathbf{p} = (p_1, p_2, \ldots)^T$ as the PMF of ${}_1T_0$, we note that $p_1 = 1/4, p_2 = 1/8$, and thereafter the following recursive relation holds:

$$ p_{k+2} = \frac{1}{2}\, p_{k+1} + \frac{1}{4}\, p_k, \quad \text{for any} \quad k \geq 1. $$

Consequently, the PMF of T_R is the same as the PMF of $X(H)$, described in (8) of Subsection 2.1, specialized to $p = 1/2$. In particular,

$$ P\{T_R = k\} = 2^{-2(k-1)} \ (1, 0) \begin{bmatrix} 2 & 1 \\ 4 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \ \text{for } k \geq 1. $$

Indeed, the PMF of T_R is $2^{-2(k-1)}a_k$, where $a_1 = 0, a_2 = 1$, and $a_{k+1} = a_k + a_{k-1}$ for $k \geq 2$, the Fibonacci sequence!

5.2.3. *Hexahedron*

Let us describe the graph of a hexahedron slightly unconventionally: Consider a hexagon whose six vertices are labeled counterclockwise as $1, 2, 3, 6, 5, 4$. Thereafter, vertices $1, 3, 5$ are edge-connected to a seventh vertex, labeled 0, below the hexagon; and vertices $2, 4, 6$ are edge-connected to an eighth vertex, labeled 7, above the hexagon. At each step, the RW goes to one of three neighbors with probability 1/3 each.

Starting from vertex 0, without loss of generality, the walk surely goes to vertex 1 (for otherwise we can rotate the hexahedron suitably). Thereafter, at each step it returns to vertex 0 with probability 1/3, or it goes to one of the vertices $2, 4, 6$ which are two edges away from vertex 0 with probability 2/3. In the latter case, in the next step it goes to vertex 7 with probability 1/3 (and then it surely returns to one of the vertices $2, 4, 6$), or it goes to one of the vertices $1, 3, 5$, which are adjacent to vertex 0, with probability 2/3. In these latter cases, holding vertices 0 and 7 fixed, we can rotate the hexahedron to ensure that the walk is either at vertex 1 or at vertex 2. Thus, we have $T_R = 1 + {}_1T_0$, where

$$ {}_1T_0 = \begin{cases} 1 & \text{w.p. } 1/3 \\ 1 + {}_2T_0 & \text{w.p. } 2/3 \end{cases} ; \quad {}_2T_0 = \begin{cases} 2 + {}_2\bar{T}_0 & \text{w.p. } 1/3 \\ 1 + {}_1T_0 & \text{w.p. } 2/3 \end{cases} $$

where ${}_2\bar{T}_0$ has the same distribution as ${}_2T_0$. Writing $\mathbf{p} = (p_1, p_2, \ldots)^T$ as the PMF of ${}_1T_0$, and $\mathbf{q} = (q_1, q_2, \ldots)^T$ as the PMF of ${}_2T_0$, we note that $p_1 = 1/3, q_1 = 0$; and defining $q_0 = 0$, the following recursive relations hold:

$$p_{k+1} = \frac{2}{3}\, q_k \quad \text{and} \quad q_{k+1} = \frac{2}{3}\, p_k + \frac{1}{3}\, q_{k-1}, \quad \text{for any} \quad k \geq 1.$$

We rewrite these recursive relations in matrix notation and iterate to obtain

$$\begin{pmatrix} p_{k+1} \\ q_{k+1} \\ q_k \end{pmatrix} = \begin{bmatrix} 0 & 2/3 & 0 \\ 2/3 & 0 & 1/3 \\ 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} p_k \\ q_k \\ q_{k-1} \end{pmatrix} = \ldots = \begin{bmatrix} 0 & 2/3 & 0 \\ 2/3 & 0 & 1/3 \\ 0 & 1 & 0 \end{bmatrix}^k \begin{pmatrix} 1/3 \\ 0 \\ 0 \end{pmatrix}.$$

The solution is formally similar to the geometric(p) distribution. To verify the total probability, one can check that

$$\begin{bmatrix} 1 & -2/3 & 0 \\ -2/3 & 1 & -1/3 \\ 0 & -1 & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1/3 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

In particular, the PMF $\mathbf{p}$ of ${}_1T_0$ simplifies to

$$p_k = P\{\, {}_1T_0 = k\} = 3^{-k} \;\; (1,0,0) \;\; \begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 1 \\ 0 & 3 & 0 \end{bmatrix}^{k-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \text{ for } k \geq 1.$$

Thus, the PMF $\mathbf{p}$ of ${}_1T_0$ is $3^{-k}a_k$, where $a_{2m} = 0$ for all $m \geq 1$; $a_1 = 1, a_3 = 4$, and $a_{2m+1} = 7a_{2m-1}$ for all $m \geq 2$.

Finally, the PMF of T_R is $(0, \mathbf{p})^T$. That is, T_R is an *even* natural number with $P\{T_R = 2\} = 1/3, P\{T_R = 4\} = 4/27, P\{T_R = 2m+2\} = (7/9)P\{T_R = 2m\}$ for all $m \geq 2$.

6. Summary and exercises

If each trial results in a success (with probability p) or a failure (with probability $q = 1-p$), then the waiting time X until success has a geometric(p) distribution, with PMF $P\{X = x\} = q^{x-1}p$ for $x = 1, 2, 3, \ldots$. We consider several more complicated notions of success and illustrate the PMF of the waiting time until success. Specifically, the success may involve a block of outcomes—either binomial or multinomial—that are independent, or Markovian, or symmetric random walks on the vertices of a given graph. Diligent researchers may extend the examples to more general situations, which may be useful in applications involving reliability systems, communication protocols, biological sequence analysis, game strategies, and stochastic network models.

To test their comprehension, student readers may solve the following exercises:

1 Without evaluating a sequence or differentiating, find $E[X^2]$, where X follows (a) a geometric(p), or (b) a negative binomial(r, p) distribution.

2 Find the PMF of the number of tosses of a fair or a biased coin until the pattern (a) HTTHT, or (b) HTHTH occurs, or (c) either HTHT or HTTT occurs.

3 How many rolls of a fair die are necessary until: (a) any one of the six values shows up three times in a row? (b) the sequence $(1, 2, 3, 4, 5, 6)$ shows up exactly in that order? (c) the six values show up successively (but in no particular order)?

4 In an R-P-S game, the two players choose their respective options according to Markov chains with transition probability matrices P and Q, respectively.

$$P = \begin{array}{c} \begin{pmatrix} .2 & .4 & .4 \\ .3 & .3 & .4 \\ .4 & .3 & .3 \end{pmatrix} \begin{array}{l} R \\ P \\ S \end{array} \text{ A's choice} \\ \begin{array}{ccc} R & P & S \end{array} \\ \text{A's next choice} \end{array}, \quad Q = \begin{array}{c} \begin{pmatrix} .3 & .3 & .4 \\ .4 & .3 & .3 \\ .2 & .4 & .4 \end{pmatrix} \begin{array}{l} R \\ P \\ S \end{array} \text{ B's choice} \\ \begin{array}{ccc} R & P & S \end{array} \\ \text{B's next choice} \end{array}$$

Neither player knows the other player's matrix. What are the probabilities of winning for Player A, Player B, and a tie?

5 Find the PMF of the return time to the starting vertex for (a) a dodecahedron and (b) an icosahedron. See these objects and their planar graphs in Wi (no date).

Acknowledgment

Dr. Vahid Andalib gave feedback on an earlier draft, and Dr. Honglang Wang mentioned several references. I thank them both. A learned referee offered many helpful suggestions to highlight the purpose, motivation, and applications of the mathematical principles advocated in this paper and to improve the exposition.

Conflict of interest

The author does not have any financial or non-financial conflict of interest to declare for the research work included in this paper.

References

Aldous, D. J. and Fill, J. A. (2014). *Reversible Markov Chains and Random Walks on Graphs*. Unpublished monograph.

Devore, J. L. and Berk, K. N. (2007). *Modern Mathematical Statistics*. Thomson-Brooks/Cole, Belmont, CA.

Donnelly, P. (2005). How juries are fooled by statistics. TED: Ideas worth spreading. See also: https://www.ted.com/talks?q=peter

Doyle, P. G. and Snell, J. L. (1984). *Random Walks and Electric Networks*, volume 22 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC.

Dudewicz, E. J. and Mishra, S. N. (1988). *Modern Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Hoel, P. G., Port, S. C., and Stone, C. J. (1971). *Introduction to Probability Theory*. Houghton Mifflin, Boston, MA.

Li, S. R. (1980). A martingale approach to the study of occurrence of a sequence pattern in repeated experiments. *The Annals of Probability*, **8**, 1171–1176.

Maiti, S. I. and Sarkar, J. (2019). Symmetric walks on paths and cycles. *Mathematics Magazine*, **92**, 252–268.

Rohatgi, V. K. (2003). *Statistical Inference*. Dover, New York, 2 edition.

Ross, S. M. (1996). *Stochastic Processes*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2 edition.

Sarkar, J. (2006). Random walk on a polygon. In *Recent Developments in Nonparametric Inference and Probability*, volume 50 of *IMS Lecture Notes*, pages 31–43. Institute of Mathematical Statistics, Beachwood, OH.

Sarkar, J. (2020). A symmetric random walk on the vertices of a hexahedron. *Mathematics Student*, **89**, 63–85.

Wi (no date). Platonic solid. Wikipedia: The Free Encyclopedia. Online; accessed 2025-10-29.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 97-113

CPI Housing Index Methodology

Ashish Das[1*] and Praggya Das[2]

[1]*Department of Mathematics, Indian Institute of Technology Bombay, Mumbai*
[2]*Former Adviser-in-Charge, Monetary Policy Department, Reserve Bank of India, Mumbai*

Received: 09 August 2026; Revised: 16 August 2026; Accepted: 20 August 2026

Abstract

This paper evaluates the methodological framework used by Ministry of Statistics and Programme Implementation (MoSPI) to compile the House Rent Index within the Consumer Price Index (CPI), comparing the Base 2012 series with the revised Base 2024 series. Utilizing a longitudinal six-panel approach, the 2024 series expands sample sizes across both urban and rural sectors and introduces an "off-cycle" flagging system to track intra-panel rent alterations between scheduled six-month visits. The paper critically analyses MoSPI's structural shift from using a weighted geometric mean to a weighted arithmetic mean for elementary-level aggregations across different housing categories and ownership statuses. The paper concludes that while arithmetic averages are appropriate for higher-level macroeconomic groupings, lower-level elementary house rent calculations must retain a geometric mean framework to preserve robustness, stability, and scientific validity in the all-India inflation measurement.

Key words: Base year; Fixed-based index; Chain-based index; Six-panel approach; Weighted geometric mean; Weighted arithmetic mean.

AMS Subject Classifications: 62D05, 91B82

*This is an extended version of the invited talk that was based on the authors' IIT Bombay Technical Report, delivered at the 28th Annual Conference of the Society of Statistics, Computers and Applications, co-hosted by Savitribai Phule Pune University, on January 5, 2026. Ashish Das is Professor in Department of Mathematics, IIT Bombay and Praggya Das is former Adviser-in-Charge, Monetary Policy Department, and former Secretary to the Monetary Policy Committee, RBI. The views expressed in this paper are those of the authors and not of IIT Bombay.

1. Introduction

Ministry of Statistics and Programme Implementation (MoSPI) produces Consumer Price Index (CPI) at a monthly frequency. CPI provides a measure of inflation that is prescribed under statute to be used by the Reserve Bank of India (RBI) for the conduct of

Corresponding Author: Ashish Das
Email: ashish@math.iitb.ac.in

monetary policy. The Base year of the current series is 2024 = 100.

CPI covers goods (food items, non-food items) and services consumed by the households. Among services covered in the current CPI, housing services are notably important. It has a weight of 11.881 per cent in the overall CPI.

MoSPI releases CPI data separately for the rural and urban areas. A combination of the CPI-Rural and CPI-Urban indices, called CPI-Combined indices, provide the measurement of all-India CPI. For the urban households, housing services account for a much higher share – 19.778 per cent of the total urban CPI, while for the rural households, housing services account for a relatively lower share – 5.527 per cent of the total rural CPI. Combining rural and urban households, the weight for housing services in all-India CPI, averages out to 11.881 per cent.

Before implementing the 2024 Base revision, on October 30, 2025, MoSPI had come out with a "Discussion Paper on the Proposed Changes in the Housing Index Compilation Methodology in the new CPI Series". An IIT Bombay Technical Report brought out by the authors on December 18, 2025, scientifically examined the methodology of house rent index compilation in the 2012 Base and changes proposed in the Discussion Paper (DP) for the 2024 Base. It deliberated on the pros and cons of the proposed changes by MoSPI in their DP.

After the release of the Technical Report, formal / informal discussions took place, on the Report, with domain-area experts, that included, officials in MoSPI, RBI, Ministry of Consumer Affairs, Food & Public Distribution, Labour Bureau, several market economists and analysts, inter alia. In this follow-up article, we provide some new insights into the CPI Housing Index Methodology with Base (2024 = 100).

Housing index in CPI is a special case. For the measurement of house rental prices in CPI, data is collected from a fixed set of rented dwellings from each of the urban and rural markets. This fixed set, or full-stock, of selected dwellings in the sample are divided into six panels. To optimise resources, longitudinal surveys are conducted where rents from each panel of dwellings are measured at six-monthly intervals. This unique method of house rent measurement was formally adopted for the housing CPI (Base 2012 = 100) and is unlike the method of price collection adopted for any other item in CPI. In their Report of the Group of Technical Advisory Committee on Statistics of Prices and Cost of Living, 2014, MoSPI mentions that "this arrangement has been made to reduce the workload of enumerators in collection of information and moreover, **rent does not change very frequently** like prices of cereals, vegetables, fruits, etc.".

The main objective of collecting data on house rent, from a fixed sample of dwellings, is to measure pure changes in rents over a period of time. The sample dwellings are selected in such a way that they represent various categories with different number of living rooms. As the six-monthly panel survey approach is adopted, the dwellings visited in the month of January for collection of rental information are visited again in the month of July. Similarly, dwellings visited in February are visited again in August, and so on.

The staggered rental price collection methodology by MoSPI is also in accordance with the international practice for house rent measurement. The 'CPI Manual: Concepts and Methods 2020', and the 'Consumer Price Index: Design (Handbook of Methods)' showcases

the international (World Bank and BLS) practice of rental housing data using a panel design. Their six-panel approach indicates that:

"*... Each month, a panel is priced, with all six panels priced twice a year: panel 1 is collected in January and July, and panel 2 is collected in February and August, and so on.* ***Every month, we collect rent prices and other information for one panel and the 6-month price ratio is computed (the current rent divided by the rent 6 months ago) for each unit in the panel.***". (See, 'Consumer Price Index: Design' (Handbook of Methods) US BLS, Last Modified Date: January 30, 2025)

Section 2 discusses the CPI 2012 series six-panel approach adopted by MoSPI for housing index computation. In contrast, Section 3 examines the modified six-panel housing index compilation methodology that has been implemented for the CPI 2024 series. Section 4 provides the similarity and distinctions in the method of housing index computation for the two series. Section 5 takes up the issue of the use of the weighted arithmetic mean as against the weighted geometric mean, while Section 6 discuses its impact on the transitivity. Finally, Section 7 gives the concluding remarks.

2. The approach of housing index compilation for base 2012

This Section recapitulates the housing index methodology used for CPI Base 2012. Unlike housing CPI 2024 series, the index for housing CPI (Base 2012 = 100) was limited to urban areas. Housing had 21.67% expenditure share for urban CPI and 10.07% expenditure share at All India level in the Base 2012 CPI series. A full-stock of rented dwellings then was about 13,368, and every month the rental prices were collected from about 2228 dwellings. Thus, in the six-panel approach, there were six sets of 2228 dwellings and each set constituting a panel. Every year, the full-stock of rented dwellings were visited such that panel-1 through panel-6 were visited from January to June, respectively, and again visited during the period July to December for canvassing the house rent schedule. While providing the six-panel coverage of the dwellings for the Base year 2012, for ease of illustration, we consider 6 panels of 2000 dwellings each in Figure 1.

The index for CPI (Base 2012 = 100) was released from January 2013 onwards. Since in each month only one-sixth of sample dwellings were visited, January 2013 had one-sixth of the rental data available. February 2013 had two-sixth of rental data available, and so on. **The full rental data of all sample dwellings was available for the first time in June 2013.** Since dwellings are visited once in six months for canvassing house rent schedule, the dwellings covered in the month of January for collection of rental prices are visited again in the month of July. Similarly, dwellings covered in February were surveyed again in August, and so on. As done for the Base year 2012, the six-panel coverage of the dwellings for the year 2013 is provided in Figure 2.

There are few other dimensions of a dwelling in a sample, the actual computation of house rent index is very nuanced. Actual House rent index computation involves the following steps:

- First, the category of dwellings (1-, 2-, 3- and $\geq$ 4-room) are accounted for. For each State and Union Territory (UT) and each category, geometric mean of rent relatives is separately worked out using the full set of dwellings in the rolling sample of six months.

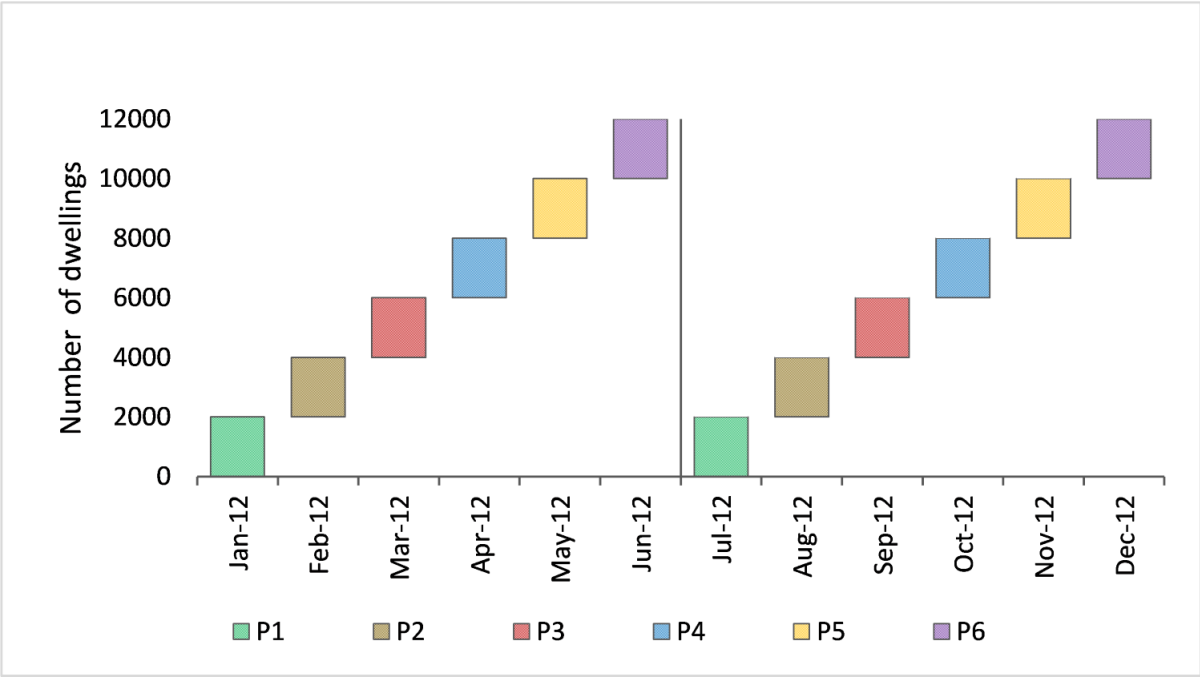


Figure 1: Monthly dwelling visits for rent collection

Note:
1. Pi pertains to dwellings in Panel i; i = 1, 2, ..., 6
2. Pi dwellings are visited in scheduled cycle
3. The slanted line blocks represent fresh visits during the month

- Second, the shares (weights) of each category of dwellings within owner-occupied and rented houses (as per the results of Housing Condition Survey NSS 69th Round) are taken. Then weighted geometric mean of the rent relatives are separately worked out for owner-occupied and rented dwellings for each State/UT. Since the rents of owner-occupied dwellings are taken by rental equivalence approach, the rent relative of rented accommodation is applied to the respective category of owner-occupied dwellings also.

- Third, the overall rent relative for a State/UT is worked out using weighted geometric mean of owner-occupied and rented dwelling. The weights of owner-occupied and rented dwellings are computed from the result of household consumption expenditure surveys (HCES) 2011-12.

- Fourth, State/UT wise house rent index is arrived at by multiplying the above rent relative to (i) 100 (in case of fixed-based index), and (ii) house rent index 6-month ago (in case of chain-based index).

- Lastly, the State/UT level house rent indices are aggregated, for arriving at the all-India index, using state-wise housing weights from the HCES. At this stage weighted arithmetic mean is used.

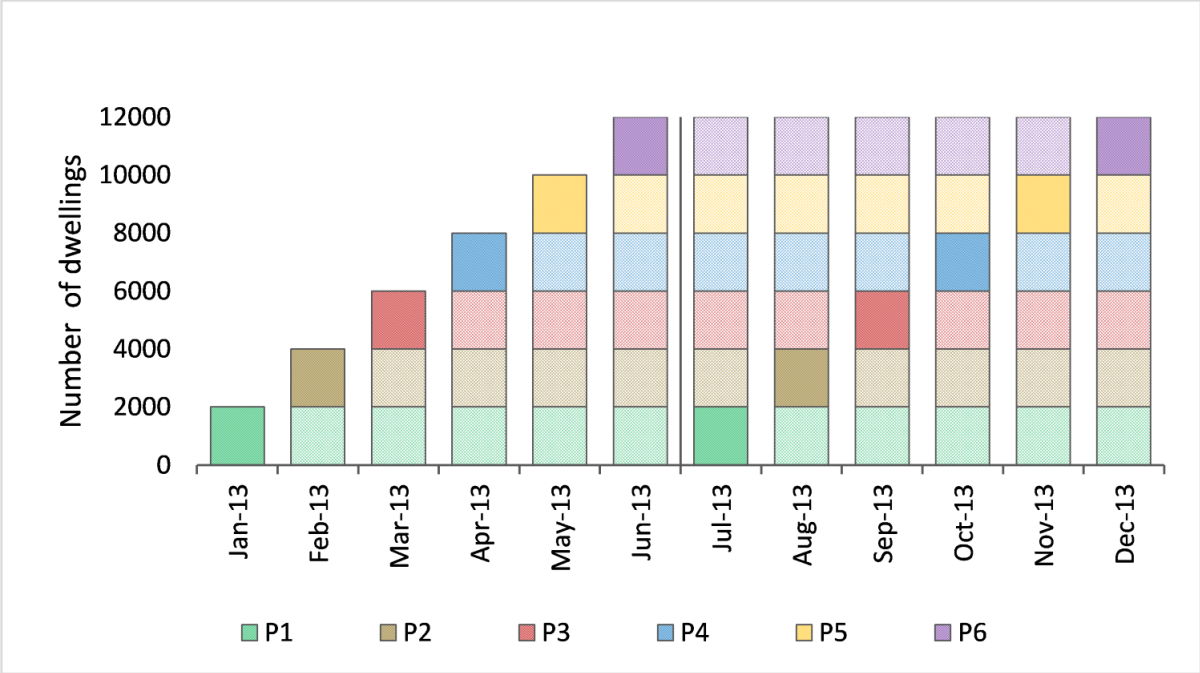


Figure 2: Rental data used for monthly index compilation

Note:

1. Pi pertains to dwellings in Panel i; i = 1, 2, ..., 6
2. Pi dwellings are visited in scheduled cycle
3. The slanted line blocks represent fresh visits during the month
4. The cross-line blocks represent the earlier visits for the panel

Computation of Index (Base 2012) – House Rent Index for a State/UT

(a) The Base rent is defined as follows: For a dwelling, if $r_1, r_2, \ldots, r_{12}$ are the rents collected in the Base year 2012 for January, February, ..., December, respectively, then the Base rent for the year 2012, as explained in the DP, is

- for January and July, $(r_1 + r_7)/2$;
- for February and August, $(r_2 + r_8)/2$; and so on; and
- for June and December, it will be $(r_6 + r_1 2)/2$.

(b) Based on the full rental data of all sample dwellings, effective June 2013, using the (weighted) geometric mean (GM) of the rent relatives, Index for the months June 2013 to November 2013 using fixed-based method is:

Weighted GM of {(Current month rent) / (Corresponding Base rent)} × 100.

(c) Based on the full rental data of all sample dwellings, effective December 2013, using the (weighted) GM of the rent relatives, Index for the months December 2013 onwards using chain-based method (6-monthly chain-based index) is:

Weighted GM of {(Current month rent) / (Rent 6-month ago)} × Index 6-month ago.

From December 2013 onwards, house rent index was compiled using a chain-based method. Under this method, current month rent was compared with the rent six months ago. Thus, the index for December 2013 is obtained by first taking the (weighted) GM of the December 2013 over June 2013 rent relatives and then multiplying by the June 2013 index. Similarly, the index for January 2014 is obtained by first taking the (weighted) GM of the January 2014 to July 2013 rent relatives and then multiplying by the July 2013 index. And so on.

3. The approach of housing index compilation for base 2024

In CPI 2024 series, the methodology for compiling the House Rent Index was comprehensively revised by MoSPI to enhance robustness, consistency, and representativeness. Unlike the CPI 2012 series, leveraging house rent and imputed rent data available from HCES 2023-24, the sample constitutes a total of 19039 dwellings (15715 in urban and 3324 in rural) for rent collection in Base (2024 = 100).

The DP indicated that the monthly rentals from the entire set of sampled dwellings would be adopted as against the panel approach. More specifically, MoSPI proposed that it will measure the rental prices from the entire set of sampled dwellings every month, instead of visiting these households once every six months. The DP mentions:

"**Monthly Rent Collection**: Rent data will be collected monthly, replacing the current practice of collecting rent every six months.

Expanded Sample Coverage: Instead of limiting rent collection to one-sixth of the sample, the new approach proposes collecting rent data from all selected dwellings each month.".

The IIT Bombay Technical Report had strong reservations to the above documented method where "Rent data will be collected monthly, replacing the current practice of collecting rent every six months" and "Instead of limiting rent collection to one-sixth of the sample, the new approach proposes collecting rent data from all selected dwellings each month".

Based on interactions after the release of the Technical Report, it transpires that MoSPI would continue to follow the six-panel coverage of the dwellings, *i.e.*, all dwellings in Panel 1 are visited in January and July; all dwellings in Panel 2 are visited in February and August; and so on; all dwellings in Panel 6 are visited in June and December.

However, a nuance has been added. While visiting the dwellings in a panel, in addition to the rental information, information is sought on the future months (in the subsequent five months) when the rent is expected to change. If the rent is expected to change for the sample dwelling, in these months, the dwelling will be flagged for revisit that month. All such dwellings are visited again in the flagged months in addition to their scheduled panel visit. This is an improvised way of data collection, by visiting few dwellings from each of the five panels in addition to visiting all the dwellings in the panel marked for the month (Figure 3 for Base year and Figure 4 for following years). This would lead to rental data collection every month from one-sixth of the full-stock plus additional data collected (say, 5

The Expert Group Report on Comprehensive Updation of Consumer Price Index (January, 2026), MoSPI, Government of India mentions "*In the new series, visit to the*

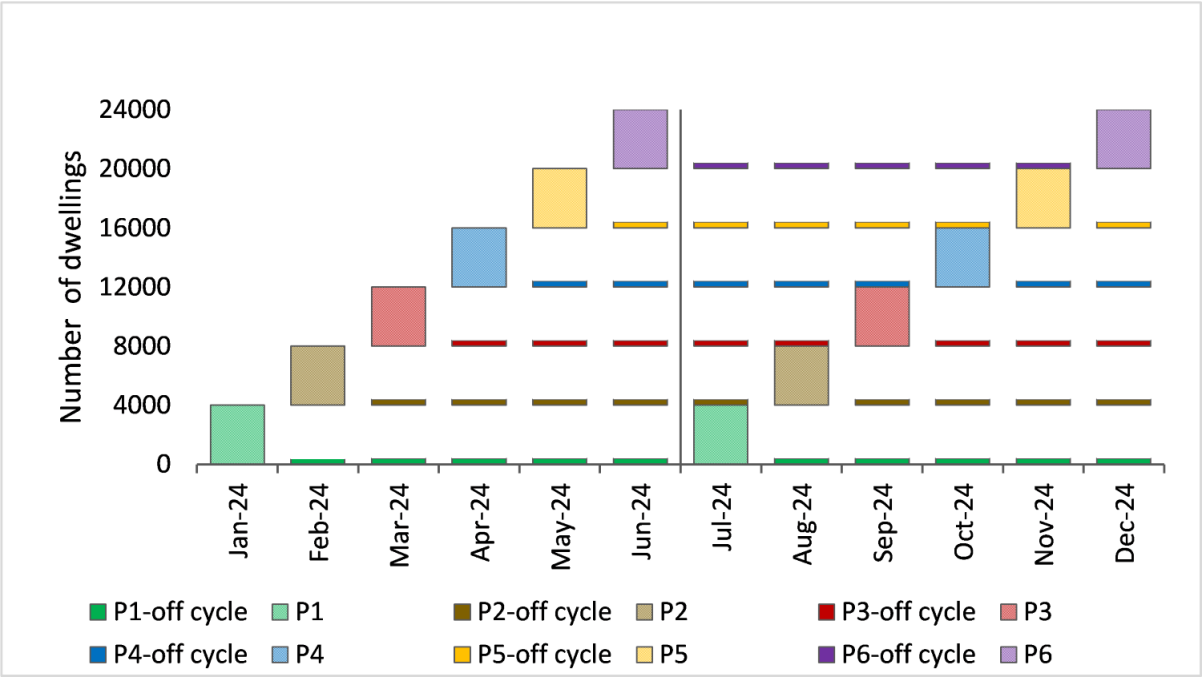


Figure 3: Monthly dwelling visits for rent collection

Note:
1. Pi pertains to dwellings of panel i; i = 1, 2, ..., 6
2. Pi dwellings are visited in scheduled cycle; Pi-off cycle are the additional dwellings visited
3. The slanted line blocks represent fresh visits during the month
4. The solid colour blocks represent the number of dwellings visited during off cycles

dwellings by the enumerator has been planned in such a way that, rent data will be available for every month. All the dwelling are distributed among 6 months and enumerator would necessarily visit each dwelling every six months. In addition, while visiting a dwelling an additional information will be collected from the respondent about the next expected date of rent revision and will flag the dwelling. Generally there is formal or informal agreement between the owner and the tenants about the period after which the rent is to be revised. If the month in which the revision in the rent is expected, falls between the two scheduled visits, the enumerator will visit the dwelling in that month also to collect the revised rent. Thus, in a month in addition to 1/6th of the dwellings additional number of dwellings which have been flagged by the enumerators will be covered. This approach would ensure that monthly data that is captured would accurately reflect changes and also reduce the need for mandatory monthly visits by the enumerator. This change addresses representativeness issues inherent in panel-based systems, particularly in markets with limited availability of rented dwellings outside major urban centres."

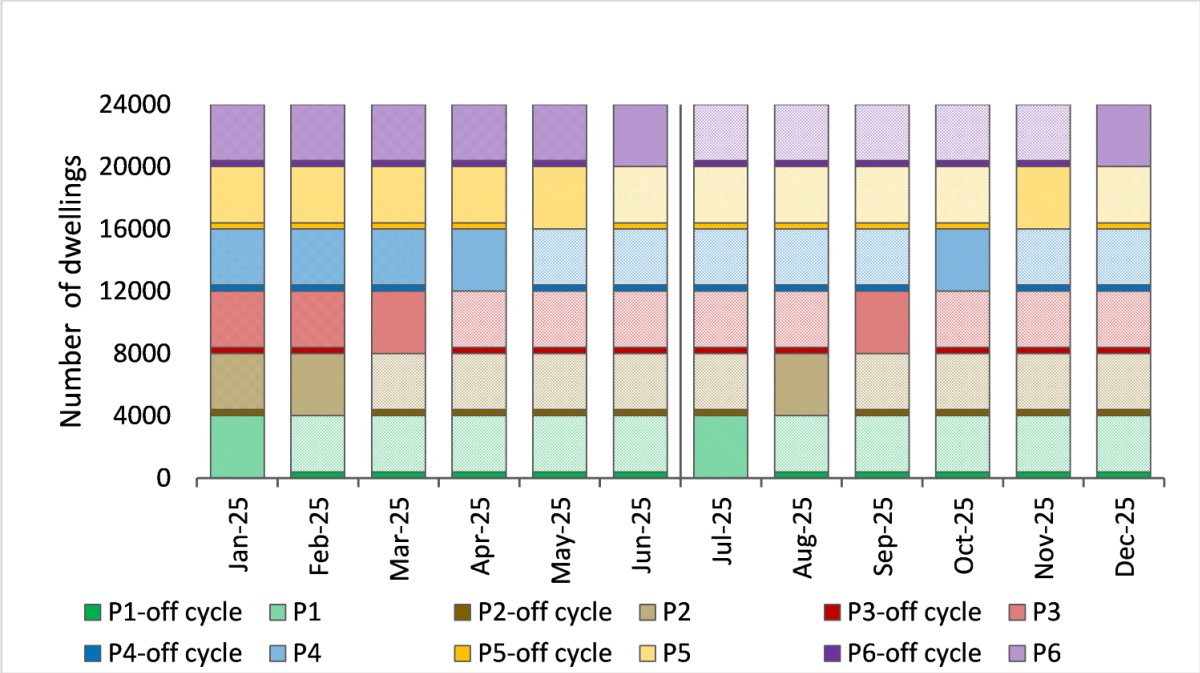


Figure 4: Rental data used for monthly index compilation

Notes:
1. Pi pertains to dwellings of panel i; i = 1, 2, ..., 6
2. Pi dwellings are visited in scheduled cycle; Pi-off cycle are additional dwellings visited
3. The light coloured blocks represent the latest dwelling visits in the base year
4. The slanted line blocks represent fresh visits during the month
5. The solid colour blocks represent the number of dwellings visited during off cycles
6. The cross line blocks represent the earlier visits for the panel

Actual House rent index computation involves the following steps:

- Dwellings are classified into four categories based on the number of living rooms (1, 2, 3, and 4 or more). Category-wise weights were derived from Census 2011, replacing the Housing Condition Survey (NSS 69th Round) used in the CPI 2012 series.
- Compilation Methodology: Housing categories are treated as priced items, and indices are compiled category-wise at the State/UT × Sector level. Here Sector level refers to Urban and Rural levels (unlike CPI 2012 series, where only Urban level existed). Monthly rent relatives are computed as the geometric mean of rent relatives within each Category and State/UT. Category-wise indices are compiled using a chain-based (short-term Jevons) index, where the current month index is obtained by multiplying the geometric mean of rent relatives with the previous month's index. Separate indices are derived for rented and owner-occupied dwellings. For owner-occupied dwellings, the rental equivalence approach is applied. Category-wise indices are aggregated using weighted arithmetic means (unlike the CPI 2012 series where weighted geometric mean is used), with weights based on category shares for rented dwelling and owner-occupied dwellings.

- The overall State/UT × Sector level House Rent Indices are obtained by aggregating rented and owner-occupied indices using weighted arithmetic means (unlike the CPI 2012 series where weighted geometric mean is used), with weights being the expenditure shares derived from HCES 2023–24. The proposed revisions aim to produce a more stable and representative Housing Index in the new CPI 2024 series.

Computation of Index (Base 2024) – House Rent Index for a House-Category

(a) **The Base rent**: For each dwelling, the Base rent is the average of the rents for the 12 months of the Base year 2024. Here, the rent for a month is based on the rent collected by an actual visit to the dwelling in a given month or rent collected based on the last visit to the dwelling.

(b) **The fixed-based January 2025 Index**: Based on the full rental data of all sample dwellings, using the geometric mean (GM) of the rent relatives, the fixed-based Index for the month January 2025 uses the borrowed base-data method, *i.e.*,

Index Jan.25 = GM of {(Jan.25 rent) / (Base rent)} × 100.

Here, the rent for January 2025 is based on the rent collected by an actual visit to a dwelling in January 2025 and rent collected based on the last visit to a dwelling in the Base year 2024 (*i.e.*, during August-December 2024). Thus, while computing the January 2025 Index, the numerator of the rent relatives uses the actual data collected in the Base year months August-December 2024. In this sense the January 2025 Index is a borrowed base-data method.

(c) **The chain-based February-May 2025 Index**: Based on the full rental data of all sample dwellings, using the GM of the rent relatives, the chain-based Index for the month February 2025 uses the borrowed base-data method, *i.e.*,

Index Feb.25 = GM of {(Feb.25 rent) / (Jan.25 rent)} × Index Jan.25.

Here, the rent for February 2025 is based on the rent collected by an actual visit to the dwellings in February 2025 and rent collected based on the last visit to the dwellings in the Base year 2024 (*i.e.*, during September-December 2024) and in January 2025. Similarly, for the months March-May 2025.

(d) **The June 2025 onwards chain-based index**: Based on the full rental data of all sample dwellings, effective June 2025, using the GM of the rent relatives, Index for the months June 2025 onwards using chain-based method (1-monthly chain-based index) is:

GM of {(Current month rent) / (Rent 1-month ago)} × Index 1-month ago.

From June 2025 onwards, house rent index was compiled using a chain-based method. The rent for June 2025 is based on the rent collected by an actual visit to all the dwellings during January-June 2025. Thus, while computing the June 2025 Index, the numerator of the rent relatives uses the actual data collected during January-June 2025. Under this method, current month rent was compared with the rent 1-month ago. Thus, the index for June 2025 is obtained by first taking the GM of the June 2025 over May 2025 rent relatives

and then multiplying by the May 2025 index. Similarly, the index for July 2025 is obtained by first taking the GM of the July 2025 over June 2025 rent relatives and then multiplying by the June 2025 index. And so on.

January 2025 Category-wise index is a fixed-based index that is based on the Current month rent of January 2025, where the Current month rent of January 2025 is from **actual visits** to dwellings in January 2025 (all Panel 1 dwellings and some dwellings from Panels 2, 3, 4, 5 and 6) and from the dwellings (in Panels 2, 3, 4, 5 and 6) **not visited** in January 2025 but by **borrowing rental values collected** from the last visit of the dwellings in the Base period August-December 2024. This leads to an inherent assumption that the rents in January 2025 did not change for those dwellings that were last visited during August-December 2024. For February-May 2025, since a 1-monthly chain-based index is being used, there is a carry-over effect in the house rent index through **borrowing of the rental values collected** in the Base months.

From June 2025 onwards, a 1-monthly chain-based index is used. However, because of the equivalence of chain- and fixed-based index, we may as well consider the June 2025 index as a fixed-based index that is based on the Current month rent of June 2025, where the Current month rent of June 2025 is from **latest actual visits** to dwellings (in Panels 1, 2, 3, 4, 5 and 6, respectively) during January-June 2025. For July 2025, the Current month rent of July 2025 is used, where the Current month rent of July 2025 is from **latest actual visits** to dwellings (in Panels 2, 3, 4, 5, 6 and 1, respectively) during February-June and July 2025. From July 2025 onwards, a 1-monthly chain-based index is used. Technically, from June 2026 onwards, the housing index would not have any carry-over effect that was prevailing in the index produced for January-May 2025.

The 1-monthly chain-based index provides some convenience over the fixed-based index. To illustrate, for Base 2024 = 100, consider that from September 2026 (*i.e.*, panel number 3), because of non-availability, a 1-room dwelling (D_o) captured in March 2026 is to be replaced by a new 1-room dwelling (D_n) in the same locality. Till August 2026, index was worked out on the basis of the rent from D_o. In September 2026, D_n's rental data is captured for both August and September 2026. To work out the chain-based September 2018 index, using the new dwelling D_n, its rent relative is obtained by computing the ratio of the September 2026 to August 2026 rent, and then multiplied to the August 2026 index (that was based on the rental data from D_o). This facilitates ease of replacing an unavailable dwelling with a new dwelling without going back to the Base rent. It is noted that starting September 2018, the chain-based index (on the basis of the rental data from D_n) will still remain equivalent to the fixed-based index that is based on the rental data from D_n. This equivalence can be established by arriving at the Base rent of D_n. The Base rent of D_n will be the Base rent of D_o multiplied by an adjustment factor – the adjustment factor being the ratio of the August 2026 rent of D_n to the August 2026 rent of D_o. However, the beauty of the chain-based approach is that we don't have to go back to the derived Base rent of D_n. For details, refer to Das and Das (2025).

4. The 2012 and 2024 series of housing CPI: Similarity and distinctions

Our analytical observations on some of the similarities and distinctions between the current housing index (CPI 2024) computation methodology and the method adopted in the

CPI 2012 series are presented below.

Both the methods use the six-panel approach by dividing the dwellings into six groups and marking a group for a mandatory visit for a month and a re-visit after every six months. Housing prices do not change frequently. Often, the rental agreements are annual, and they keep the rental values constant for about a year. Thus, a rolling sample of six months, broadly captures the latest rental values of each sample household (of the full-stock).

However, a nuance has been added in the CPI 2024 series. While visiting the dwellings in a panel, in addition to the rental information, information is also sought on the future month (in the subsequent five months) when the rent is expected to change. If the rent is expected to change for the sample dwelling, in these months, the dwelling is flagged for revisit that month. All such dwellings are visited again in the flagged month in addition to the scheduled panel visit for the month. This is an improvised way of data collection, by visiting few dwellings from each of the five panels in addition to visiting all the dwellings in the panel scheduled for the month.

In the CPI 2012 series, a rent from a dwelling is captured for a month by an actual visit and in the subsequent five months it is assumed that the rent of the dwelling does not change. In other words, the contribution towards month-on-month index (September-on-August index, say) is only from dwellings in a single Panel (September-Panel). In contrast, the CPI 2024 series, depending on the response to the additional information, assumes that the rent of the dwelling does not change either in the subsequent five months, or in the subsequent four months, or in the subsequent three months, or in the subsequent two months, or in the subsequent one month. Accordingly, here the contribution towards month-on-month index (September-on-August index, say) is from dwellings in a single Panel (September-Panel) plus all dwellings in the other Panels (Panels 1, 2, 4, 5, 6) who are visited in September.

The CPI 2024 series uses a 1-monthly chain-based index while the CPI 2012 series used the 6-monthly chain-based index. Accordingly, under the equivalence of fixed-based and chain-based index, if there are no dwellings that have a change of rent in the subsequent five months since last visited for rent collection, the CPI 2024 method reduces to the same method as adopted in the CPI 2012 series, and are thus equivalent.

The distinction in the two methods, at the category level index computation, arise only from collection of additional rental data (CPI 2024 series) from some of the dwellings visited ahead of its scheduled six-monthly panel visits. The housing index methodology in CPI 2024 series is thus like "*Old Wine in a New Bottle*".

Unlike the CPI 2012 series where weighted geometric means are used, in the CPI 2024 series, for each State/UT (urban and rural), arithmetic mean of rent relatives is used throughout for combining the shares of each category of dwellings within owner-occupied and rented houses. The weighted arithmetic mean of the rent relatives is separately worked out for owner-occupied and rented dwellings for each state. The overall rent relative for a State/UT (urban and rural) is worked out using weighted *arithmetic mean* of owner-occupied and rented dwellings.

5. Weighted arithmetic mean against the weighted geometric mean

For higher level indices (arrived by aggregation of elementary indices), the CPI 2024 series use weighted arithmetic mean for house rent (in line with other items in CPI) in place of weighted geometric mean. Bringing in uniformity for arriving at higher level indices across commodities is a well thought out. However, house rental is a special commodity as it has significant weight in CPI and has multiple dimensions – it has different categories (based on number of rooms in the dwellings), and different status of dwelling (self-owned or rented).

In this context, it will be pertinent to run through some of the fundamental concepts behind compilation of elementary indices and higher level indices (See, Appendix, reproduced from Das and Das (2025)).

For CPI compilation, it is critical to differentiate between the construction of elementary indices and higher-level indices. Elementary index is the lowest level price index in CPI. It is arrived at by aggregating price relatives of priced items. The aggregation may involve more than one step – first price relatives are aggregated separately for each of the varieties/types/categories within an item. Next these aggregate price relatives are further combined using weights (for their varieties/types/categories) to the final elementary index. These steps of aggregation of price relatives across varieties/types/categories is still a step towards construction of elementary index only and should not be construed as preparation of a high-level index. The higher-level indices include grouping the elementary item indices to sub-group/group indices and further to headline indices; or grouping them to regional indices (state-wise aggregation) and further to all-India indices.

In case of house rent index, the elementary level aggregation involves multiple steps. Aggregating rent relatives across different dwelling categories (*e.g.*, 1-room, 2-room) or across ownership (owner-occupied or rented) are steps in constructing the elementary house rent index. Till such time the rent relatives are being aggregated, geometric mean should be used to arrive at elementary house rent index. Thereafter, for higher level indices (housing index by aggregating house rent index, watchman charge index, water charge index, *etc.*, or for combining state house rent indices to arrive at all-India house rent index), weighted arithmetic mean should be used. For combining rent relatives, taking arithmetic mean instead of geometric mean, sacrifices on the benefits of geometric mean to capture the true overall change in the rent-relative ratios.

For combining the four categories of rent relatives, taking arithmetic mean, instead of geometric mean, sacrifices on the benefits of geometric mean to capture the true overall change in the house rentals, measured through the rent relatives. For example, consider 1-room house rents increase by 5 per cent, while in a small part of the market, high end 3-room house-rents increase by 50 per cent. Arithmetic mean will be heavily skewed by that 50 per cent jump. It is a subtle but really a vital point. The arithmetic mean gives too much weight to items that have had extreme percentage increases. It would make the overall change in house rent seem much higher than the average consumer is truly feeling. Arithmetic mean can distort the perception of the real change. It sacrifices the accuracy and stability that the geometric mean provides. Moreover, the Report of the Group of Technical Advisory Committee on Statistics of Prices and Cost of Living (2014) also recommends that for compilation of elementary indices,

"Consistent with the international practices, Geometric Mean may be used to compile elementary indices because of the following reasons:

- *geometric mean is stable and is robust against extreme values; and*
- *average of the Price Relatives with respect to Base Prices and Ratio of the Averages of Prices with respect to Base Prices coincide in the geometric mean."*

Therefore, rather than treating house rent like any other item in CPI, a separate treatment of house rent index compilation, as was scientifically done in the CPI 2012 series should continue. Bringing in arithmetic mean in elementary index calculation, is sub-optimal and lacks the desirable property of transitivity (as explained in the next Section). After arriving at the elementary house rent indices aggregated across categories and ownership, higher level indices may be compiled through arithmetic mean.

6. Transitivity

An important property of the indices is that of it being transitive (See, Para 8.25 of the international *CPI Manual: Concepts and Methods (2020)*). Transitivity means that the chained monthly indices are identical to the corresponding direct indices. This property is important in practice, because only then elementary price indices can be calculated as chained indices that link together the month-on-month indices. The elementary house rent indices are computed using the Jevons index. Jevons index is defined as the unweighted geometric mean of the price relatives, which is identical to the ratio of the unweighted geometric mean prices, for the two periods, 0 and t, to be compared. The Jevons price index is calculated by directly comparing the prices of the two periods, 0 and t. Indices that are calculated by comparing the price of the reference period and the current period directly are referred to as direct indices. The Jevons index is transitive.

The Expert Group Report on Comprehensive Updation of Consumer Price Index (January, 2026), MoSPI, Government of India provides illustrations on the compilation of indices for Weighted and Priced items. As an example, it mentions that in the fruit's category, HCES collected expenditure data for the combined item "Pears/Nashpati". The expenditure reported under this item represented the total household expenditure on both fruits consumed during the reference period. In contrast, in the market, these fruits were sold and purchased separately. Accordingly, while "Pears/Nashpati" appeared as a single weighted item in HCES, it was mapped with multiple priced items in CPI price collection—namely, "Pears" and "Nashpati". This distinction between weighted items and priced items ensured that the CPI basket remains conceptually meaningful, while price collection continues to be operationally feasible and representative of actual consumption.

In such cases, elementary indices would be compiled separately for Pears and Nashpati. The CPI 2024 series adopts a method where the combined index for the weighted item "Pears/Nashpati" would be derived by taking the *weighted arithmetic mean* of the two elementary indices. Here, equal weights are assigned to each of the priced items, Pears and Nashpati. Such an approach of taking the *weighted arithmetic mean* (instead of a *weighted geometric mean*) fundamentally distorts the computation of the index for Base 2024 = 100 for the weighted item "Pears/Nashpati". The distortion is in the sense of loss of *transitivity* for the index computation of the weighted item "Pears/Nashpati".

For items such as "Rail fare" and "Electricity charges" *etc*, a single weighted item may correspond to multiple priced items. In such cases, priced item–wise weights are separately available from administrative data sources, such as number of passengers, revenue by fare category, or number of domestic users, *etc*. These weights are usually appropriate proxy weights based on administrative indicators such as passenger counts, usage levels, or revenue shares. Here again for the CPI 2024 series, the weighted item index being derived from the individual priced item indices by applying a *weighted arithmetic mean*, (using the available priced item–wise weights) would distort the derivation of the true index of the weighted items like "Rail fare" and "Electricity charges". Instead, a *weighted geometric mean* for the index computation (Base 2024 = 100) of the weighted items like "Rail fare" and "Electricity charges" would retain the fundamental property of transitivity for the index computation of the weighted items "Rail fare" and "Electricity charges".

On similar lines, the computation of the index for the weighted item 'House rent' for the Base 2024 = 100 provides a distorted index number. Aggregating rent relatives across different dwelling categories (*e.g.*, 1-room, 2-room) or across ownership (owner-occupied or rented) are steps in constructing the elementary house rent index. By taking a *weighted arithmetic mean* instead of the weighted geometric mean for combining the category-wise rent relatives induces a loss of transitivity for the index computation. The same is true while combining across ownership type.

7. Concluding remarks

We have focused in this Report on the changes MoSPI has undertaken on housing index compilation methodology. The Report examines each of the changes introduced in the CPI 2024 series and provides reinforcements to establish some pros and cons in the housing index compilation methodology. Importantly the Report scientifically justifies the need to continue with panel sample rent collection approach and continuation of elementary index compilation based on geometric means. Our strongly believe that bringing in arithmetic mean in elementary index calculations, will be sub-optimal in various aspects.

Data-driven decision-making would be accurate not only by collecting the correct data but also by correctly aggregating the same through precise summary statistics and measures. The sample survey-based house rent index compilation is an exercise of national importance as it contributes significantly towards inflation measurement of the country. It is our constitutional fundamental duty to *develop the scientific temper, and the spirit of inquiry and reform*, and the study is carried out in that spirit. MoSPI may like to deliberate on the suggestions made in the Report.

References

Consumer Price Index Manual: Concepts and Methods (2020). International Monetary Fund, International Labour Organization, Statistical Office of the European Union (Eurostat), United Nations Economic Commission for Europe, Organisation for Economic Co-operation and Development, The World Bank. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@stat/documents/publication/wcms_761444.pdf.

Consumer Price Index: Design (Handbook of Methods) US BLS https://www.bls.gov/opub/hom/cpi/design.htm.

Das, P. (2018a). Impact of increase in house rent allowance on CPI inflation, Mint Street Memo No. 11, Reserve Bank of India, April 23, 2018. https://www.rbi.org.in/Scripts/MSM_Mintstreetmemos11.aspx.

Das, P. (2018b). Housing services in CPI – Measurement issues, Mint Street Memo No. 15, Reserve Bank of India, November 5, 2018. https://www.rbi.org.in/Scripts/MSM_Mintstreetmemos15.aspx.

Das, P. and Das, A. (2025). *Consumer Price Index: Methodological Aspects of Measuring Housing Services*. IIT Bombay Technical Report, December 18, 2025. https://dspace.library.iitb.ac.in/jspui/handle/123456789/43734.

Expert Group Report on Comprehensive Updation of Consumer Price Index (January, 2026), Price Statistics Division, National Statistics Office, MoSPI, Government of India. https://www.mospi.gov.in/uploads/Marquee/doc-0aee0026-bf32-4213-ae4f-c5308088444e.pdf.

Monetary Policy Report (April 2018). Reserve Bank of India, April 5,2018. https://www.rbi.org.in/Scripts/PublicationsView.aspx?id=18110.

Ministry of Statistics and Programme Implementation (2025). Proposed changes in new CPI series, Government of India, November 26. https://docs.google.com/presentation/d/1DCXdzC2_hymsqs5BppGPPgBWlCnREWx_/mobilepresent?slide=id.p1.

National Statistics Office (2024). Report No.591, *Survey on Household Consumption Expenditure 2022-23*, Ministry of Statistics and Programme Implementation, Government of India, June. https://www.mospi.gov.in/sites/default/files/publication_reports/Report_591_HCES_2022-23New.pdf.

National Statistics Office (NSO) (2025). Report No.592, *Survey on Household Consumption Expenditure 2023-24*, Ministry of Statistics and Programme Implementation, Government of India, January. https://nccd.gov.in/uploads/Household_Consumption_Expenditure2023_24_d9fcc28952.pdf.

National Statistics Office (2025). Discussion paper on the proposed changes in the housing index compilation methodology in the new CPI series, Ministry of Statistics and Programme Implementation, Government of India, October 31. (See, Annex 4) https://new.mospi.gov.in/uploads/announcements/announcements_1761816278170_89e91cb4-018c-44b1-87f7-17a8e45db702_Discussion_paper_on_the_proposed_changes_in_the_Housing_index_compilation_methodology_in_the_new_CPI_series.pdf.

Report of the Group of Technical Advisory Committee on Statistics of Prices and Cost of Living (2014). Prices & Cost of Living Unit, National Accounts Division, Central Statistics Office, Ministry of Statistics and Programme Implementation, Government of India. https://cpi.mospi.gov.in/PDFile/Report_of_the_Group_of_TAC_on_SPCL_on_Base_Revision_of_CPI.pdf.

Appendix

Elementary Indices and Higher Level Indices

Consumer Price Indices (CPIs) measure the movement in prices of an aggregate basket of goods and services consumed by a typical household in a country. The goods and services items that are included for construction of a CPI are aggregated using their expenditure share obtained from the household consumption expenditure surveys (HCES).

Calculation of CPI is usually done in two stages – it starts with compilation of elementary index for each item and then the elementary indices are aggregated to higher level indices using the item's expenditure weights in HCES.

Elementary Price Indices

Elementary index is called elementary since it is the lowest level price index and corresponds to a **weighted item** in CPI. Weighted items are all the items that get weight in CPI and for which indices are released by National Statistics Offices. Expenditure weights are usually not available below the elementary aggregate level. In case of India, the CPI (Base 2012 = 100) series consists of 299 weighted items and all India indices for these items are released every month by National Statistics Office (NSO).

To compile the 299 item indices, NSO collects prices for many more items than the 299 items. These are called **priced items**. Priced items reflect different varieties of a weighted item (*e.g.*, 'rice' may include priced items such as basmati, parmal, ponni), or different types of a weighted item (*e.g.*, 'other vegetables' may include raw papaya, sweet potato, *etc.*) or different categories of a weighted item (*e.g.* 'house rent' includes, 1-room, 2-room, 3-room or $\geq$ 4-room dwellings; or owner-occupied and rented dwellings).

The priced items are aggregated to elementary item indices and the aggregation may involve more than one step. As an example, we take the case of rice and consider that it has three varieties. To work out the elementary index of rice, first the geometric mean of price relatives for each variety of rice (priced items) from different markets is worked out. The geometric means of these three varieties of rice are then aggregated as a weighted mean, if weights are available; or as a simple mean, if weights are not available. The computation of elementary index of rice is, thus, a two-step process. The same analogy of elementary index compilation for different varieties of priced items can be extended to different types and different categories of priced items.

For a weighted item of CPI, which has different varieties, if the expenditure weights of each of these varieties are not available, the choice of the averaging method can be very important. The larger the variation in the individual prices, the larger the difference among the standard averaging methods. Though both arithmetic and geometric averaging can be used, the international 'CPI Manual: Concepts and Methods 2020' recommends the use of geometric average formulas. (Para 1.145 of the international CPI Manual: Concepts and Methods (2020))

All the steps of aggregation of price relatives to arrive at the elementary index of weighted items are part of elementary index compilation and should not be construed as preparation of a higher-level index. Also, till the elementary index is prepared, the aggrega-

tion deals with price relatives only and the method of aggregation should remain the same – geometric mean, as recommended by the international 'CPI Manual: Concepts and Methods 2020'.

Higher Level Indices

Elementary indices provide the basic building block for higher level indices and are the starting point for calculating CPI. Working upward from the elementary price indices, **all indices above the elementary aggregate level are higher-level indices** that can be calculated from the elementary price indices using the elementary aggregate expenditure as weights. (Para 8.10 of the international CPI Manual: Concepts and Methods (2020))

The higher level aggregates include grouping the elementary item indices to sub-group/group indices (vertical aggregation); or to regional indices (horizontal aggregation).

Higher level indices are calculated from the elementary price indices using the household expenditure as weights and using weighted arithmetic mean for aggregation. Usually, for higher level indices, Laspeyres Index or Young Index formula of weighing the elementary indices is used.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 115-128

Performance Assessment of Statistical, Machine Learning and Hybrid Forecasting Approaches for Pearl Millet Production in the Jammu Division

Divyam Sharma, Manish Sharma and Nishant Jasrotia

Division of Statistics and Computer Science, SKUAST-Jammu

Received: 16 July 2026; Revised: 19 August 2026; Accepted: 23 August 2026

Abstract

Pearl millet (Pennisetum glaucum) is an important cereal crop contributing to food and nutritional security, necessitating accurate production forecasting for effective agricultural planning. The present study evaluated and compared the forecasting performance of statistical and machine learning-based time series models for pearl millet production in the Jammu division using secondary data for the period 1966–2024. Simple Exponential Smoothing (SES), Double Exponential Smoothing (DES), Autoregressive Integrated Moving Average (ARIMA), Nonlinear Support Vector Regression (NLSVR), and Hybrid ARIMA-SVR models were developed and evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Diebold–Mariano test, and the Chi-square goodness-of-fit test. The Brock–Dechert–Scheinkman (BDS) test rejected the null hypothesis of independent and identically distributed residuals, indicating significant nonlinear dependence in the ARIMA residuals and supporting the use of machine learning models. Among all models, NLSVR recorded the lowest testing RMSE (0.42) and MAE (0.31) and adequately fitted the historical data and hence is the most suitable model for forecasting pearl millet production in the Jammu division.

Key words: Pearl millet; Time series forecasting; ARIMA; NLSVR; Hybrid ARIMA-SVR; Machine learning.

AMS Subject Classifications: 62D05, 91B82

1. Introduction

Agriculture is the backbone of the Indian economy and according to the economic survey 2024-2025, it contributes approximately 17 to 18% to the country's Gross Value Added (GVA) and providing livelihoods to about 46.1% of the country's workforce, thereby making a significant contribution to national economic growth, rural development, and food security (Mehmood, 2025). India possesses diverse agro-climatic conditions, which facilitate the cultivation of a wide variety of crops. Since independence, the agricultural sector has

Corresponding Author: Manish Sharma
Email: Manshstat@gmail.com

made remarkable progress in terms of area, production, and productivity of several crops. Among cereal crops, pearl millet (Pennisetum glaucum), commonly known as bajra, ranks sixth globally and fourth after rice, wheat, and maize in India (Rani *et al.*, 2024). It is one of the major food crops cultivated in the arid and semi-arid regions of the country and is valued for its drought tolerance and nutritional quality. Pearl millet is believed to have originated in Africa and later spread to India. It is widely cultivated in countries such as India, Nigeria, Sudan, and Saudi Arabia and is commonly known as bulrush millet, bajra, or cattail millet. The grains are rich in nutrients, containing 9 to 15 per cent protein, about 5 per cent fat, and 2 to 7 per cent minerals, besides being good sources of vitamins A and B, thiamine, and riboflavin (Pal *et al.*, 1996).

Pearl millet is mainly cultivated during the kharif season in India. However, summer cultivation under irrigated conditions generally provides higher yields and better grain quality than kharif cultivation (Reddy *et al.*, 2013). Summer cultivation also experiences lower disease incidence due to the disease escape mechanism. Nevertheless, kharif pearl millet occupies a larger area because rainfall substantially reduces irrigation costs. Despite its importance, prolonged dry spells and adverse climatic conditions often affect crop productivity. Keeping in view the importance of millet, the United Nations General Assembly at its 75th session in March 2021, at the behest of the Government of India, declared 2023 the International Year of Millets (IYM 2023), with the aim to create awareness about health and nutritional benefits of millets as also to increase production and consumption of millets. During 2024–25, pearl millet in the Jammu division recorded a production of 8.69 thousand tonnes (Directorate of Economics and Statistics, Jammu and Kashmir, 2025). Although bajra is the principal millet crop grown in the Union Territory of Jammu and Kashmir, its productivity remains relatively low, necessitating the adoption of improved cultivars and appropriate package of practices to ensure sustainable cultivation.

Considering the economic, nutritional, and climatic significance of pearl millet, it is essential to study its production behaviour over time and assess future prospects. Time series analysis provides an effective framework for understanding historical trends and generating reliable forecasts. In this context, various statistical and machine learning approaches such as exponential smoothing techniques Simple Exponential Smoothing (SES) and Double Exponential Smoothing (DES), Autoregressive Integrated Moving Average (ARIMA) models, Nonlinear Support Vector Regression (NLSVR), and hybrid ARIMA–SVR models have been widely adopted to capture both linear and nonlinear patterns in agricultural time series data. Exponential smoothing methods are simple yet powerful tools for short-term forecasting, while ARIMA models are effective in modelling linear dependencies in time series data. However, agricultural data often exhibit nonlinear behaviour due to climatic variability and socio-economic factors. To address this limitation, machine learning techniques such as NLSVR are employed. Furthermore, hybrid models like ARIMA-SVR combine the strengths of both statistical and machine learning approaches, leading to improved forecasting accuracy. Therefore, the present study was undertaken to evaluate and compare the forecasting performance of exponential smoothing, ARIMA, NLSVR, and hybrid ARIMA-SVR models for pearl millet production in the Jammu division and to identify the most appropriate model for accurate future forecasting.

2. Material and methods

Data: The study was carried out for the Jammu region. The secondary data on pearl millet production and area (in hectares) for a period of 59 years (1966-2024) were collected from various authentic sources, including the Directorate of Economics and Statistics, Jammu and Kashmir, and the Indian Institute of Millets Research.

Statistical time series models:

(i) Exponential smoothing: Exponential smoothing is a widely used time series forecasting technique that generates forecasts by assigning exponentially decreasing weights to past observations, thereby placing greater emphasis on more recent values. Depending on the characteristics of the time series, Simple Exponential Smoothing (SES) and Holt's Linear Trend (Double Exponential Smoothing) methods may be employed. SES is appropriate for series exhibiting no systematic trend or seasonality, whereas Holt's method extends SES by incorporating a linear trend component (Holt, 1957). The SES forecasting equation is given by

$$F_{t+1} = F_t + \alpha(Y_t F_t)$$

where α is the smoothing constant, taking values between 0 and 1, and F_t is the forecast for Y_t.

For Holt's linear trend method, the level and trend components are estimated as

Level equation: $l_t = \alpha Y_t + (1-\alpha)(l_{t-1} + b_{t-1})$
Trend equation: $b_t = \gamma(l_t - l_{t-1}) + (1-\gamma)b_{t-1}$
Forecast equations: $Y_{t+1} = l_t + hb_t$

where, l_t denotes an estimate of the level of the series at time t, bt denotes an estimate of the trend (slope) of the series at time t, α is the smoothing parameter for the level and γ is the smoothing parameter for the trend which range from 0 to 1.The last estimated level plus h times the last estimated trend value is the h-step-ahead forecast. Forecasts thus represent a linear function of h.

(ii) Auto regressive integrated moving average: ARIMA (Auto regressive integrated moving average) models are widely used for modelling and forecasting univariate time series data, particularly when non-stationary series can be transformed into stationary series through differencing. An ARIMA model is denoted as ARIMA (p, d, q), where p represents the auto regressive order, d denotes the degree of differencing required to achieve stationarity, and q represents the moving average order. The general form of an ARIMA model is expressed as:

$$Y_t = \sum_{i=1}^{p} \beta_i Y_{t-i} + e_t + \sum_{j=1}^{q} \phi_j e_{t-j}$$

where Y_t is the observed value at time t, α is the intercept, β_i are the autoregressive coefficients, ϕ_j are the moving average coefficients, and e_t is the random error term (Box and Jenkins, 1976).

Machine learning time series models

(i) Non linear support vector regression (NLSVR) model: NLSVR primarily

converts the initial input space into a feature space with higher dimensions. Subsequently, a linear regression model is built within it, effectively representing non-linear regression in the original space. In the context of a dataset represented by $Z = \{X_i Y_i\}_{i=1}^{N}$, where the input vector X_i belongs to the n-dimensional real space, Y_i represents the scalar output, and N denotes the size of Z, the general equation of NLSVR can be expressed as follows:

$$f(x) = W^T \phi(x) + b$$

where w represents the weight vector, $\phi(x)$ stands for the non-linear mapping function, b signifies the bias term, and superscript T indicates the transpose operation. The data is used to estimate the coefficients w and b through the minimization of a regularized risk function:

$$R(\theta) = \tfrac{1}{2}\|W\|^2 + C\left[\tfrac{1}{N}\sum_{i=1}^{N} L_E(Y_i, f(X_i))\right].$$

The above equation comprises two elements: the first is a regularized component represented as half times the norm of w squared, while the second component is referred to as the empirical error, denoted as $\frac{1}{N}\sum_{i=1}^{N} L_E(Y_i, f(X_i))$. The regularized risk function effectively balances the optimization of both of these components simultaneously, thus preventing both underfitting and overfitting issues of the model.

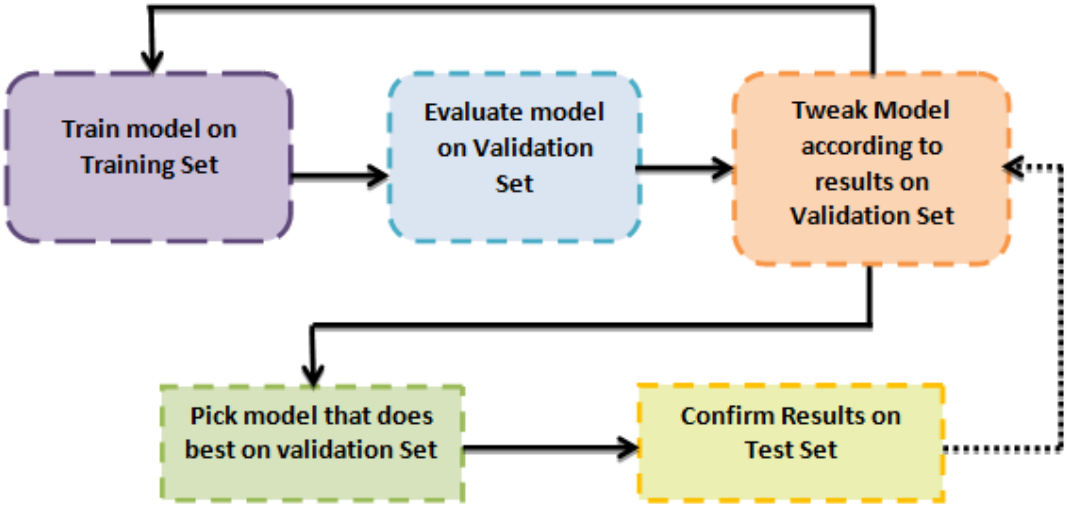


Figure 1: Standard machine learning workflow using training, validation, and test datasets

(ii) Hybrid models: Hybrid time series models combine statistical and machine learning techniques to improve forecasting accuracy by capturing both linear and nonlinear patterns in the data. In this study, an ARIMA–SVR hybrid model was used for forecasting pearl millet production in the Jammu region. First, an ARIMA model is fitted to the original series to capture the linear structure, and residuals are obtained as:

$$e_t = Y_t - \hat{Y}_t^{(ARIMA)}.$$

Since residuals may contain nonlinear information, Support Vector Regression (SVR) is applied to model them. The final hybrid forecast is generated by combining both components:

$$\hat{Y}_t^{(Hybrid)} = \hat{Y}_t^{(ARIMA)} + \hat{e}_t^{(SVR)}.$$

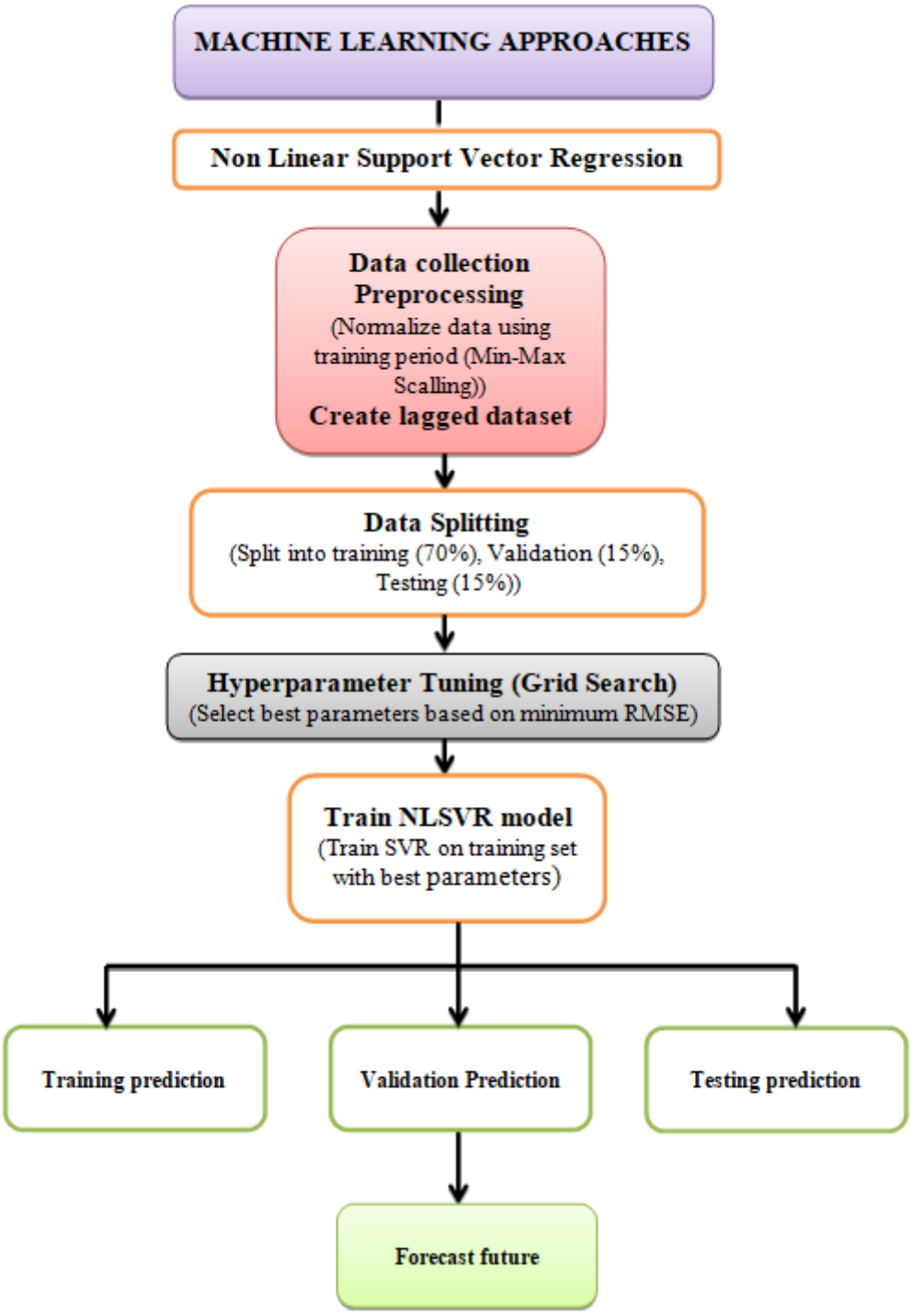


Figure 2: Standard workflow of the NLSVR-based machine learning approach

Train test validation split

The dataset was divided into three subsets: training, validation, and test datasets. The training set (70% of the data) was used to develop and fit the predictive model by learning underlying patterns from the data. Model performance was subsequently evaluated using the validation set (15% of the data), which facilitated hyperparameter tuning and optimization through iterative adjustments. Based on validation results, the best-performing model was selected for further analysis. Finally, the selected model was evaluated on the independent test set (15% of the data), which was not involved in model development, to obtain an unbiased estimate of predictive accuracy, robustness, and generalization capability on unseen data.

Model adequacy

To evaluate model performance, RMSE and MAE were computed for all forecasting models.. Lower values of these measures indicated better model performance and prediction accuracy. The RMSE and MAE is given by the formulas:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{(n-2)}}$$

where, Y_t is the Actual value, $\hat{Y}_t$ is the Predicted value and n is the total number of observations

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t|$$

where, n is the number of error variables and $|Y_t - \hat{Y}_t|$ is the absolute error.

3. Results and discussion

The descriptive statistics of pearl millet production during 1966-2024 revealed a mean production of 8.15 thousand tonnes, with values ranging from 3.70 to 13.59 thousand tonnes. The standard deviation (2.22 thousand tonnes) and coefficient of variation (27.27%) indicated moderate variability in production over the study period. The skewness (0.16) and kurtosis (2.31) values suggested an approximately symmetric and slightly platykurtic distribution. The Durbin–Watson statistic (DW) is 1.10 indicated the presence of positive autocorrelation in the series, justifying the application of time-series models for capturing temporal dependence and generating reliable forecasts.

Table 1: Descriptive statistics of pearl millet production from 1966-2024

Variable	N	Mean	Max	Min	SD	CV (%)	Skewness	Kurtosis	Durbin –Watson test
Production (000tonnes)	59	8.15	13.59	3.7	2.22	27.27	0.16	2.31	1.1

As shown in figure 3, pearl millet production did not remain constant from 1966 to 2024. Production increased in some years and decreased in others, showing noticeable fluctuations over time. These ups and downs indicate that production is influenced by various

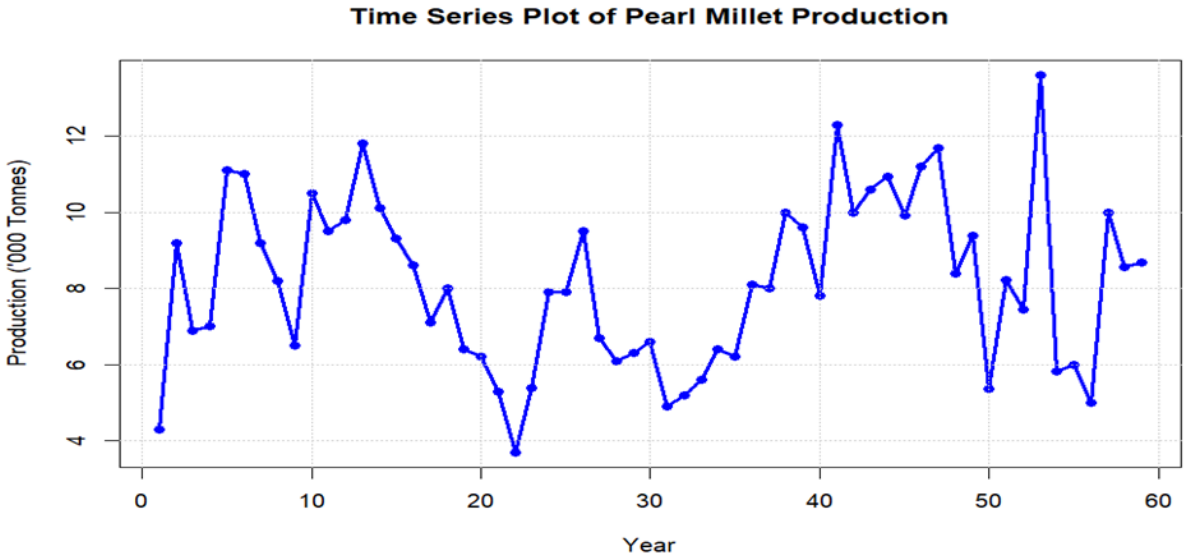


Figure 3: Graphical representation showing fluctuations in data of pearl millet crop for the period 1966-2024

factors and follows a time-dependent pattern, making it suitable for forecasting using time-series models.

Table 2: Simple exponential smoothing and double exponential smoothing parameters

Models	α	β
Simple exponential smoothing	0.5	-
Double exponential smoothing	0.1	0.6

Simple Exponential Smoothing (SES) and Double Exponential Smoothing (DES) models were employed to forecast pearl millet production. The smoothing parameters were optimized through a grid-search procedure using the training and validation datasets. The optimal parameter for SES was $\alpha = 0.5$, indicating that equal weight was assigned to recent and past observations. For DES, the optimal values were $\alpha = 0.1$ and $\beta = 0.6$, suggesting that the model emphasized long-term historical information while effectively capturing trend variations in the data. These parameter values were selected based on the minimum RMSE and MAE.

As shown in figure 4, the ACF plot exhibits a gradual decline in autocorrelation across lags, indicating the presence of temporal dependence in pearl millet production data. The PACF plot shows significant spikes at the initial lags, suggesting that autoregressive components may be present and that the series is suitable for time-series modelling.

The ADF test results in Table 3 showed that the original pearl millet production series was non-stationary, whereas the first-differenced series became stationary.

As shown in Table 4, four ARIMA models were evaluated using RMSE and MAE

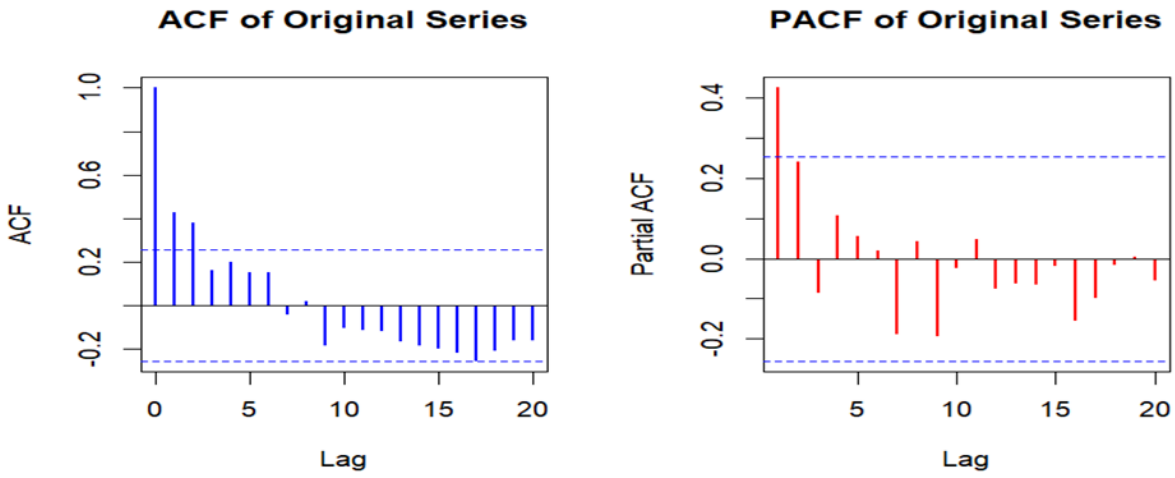


Figure 4: Graphical representation of ACF and PACF of original series for production in J&K

Table 3: Stationarity test results for pearl millet production under J&K through ADF test

	Original series			First differenced series		
	Statistics	Lag	p-value	Statistics	Lag	p-value
ADF Test	-2.03	3	0.55	-4.04	3	0.01

values across training, validation, and testing datasets. Among the candidate models, ARIMA(1,1,1) exhibited the lowest validation RMSE (1.15) and MAE (0.91), indicating superior predictive performance. Therefore, ARIMA(1,1,1) was selected as the most suitable model for forecasting pearl millet production in Jammu and Kashmir.

The parameter estimates and diagnostic statistics reported in Table 5 demonstrated that the model parameters were significant and that the residuals behaved as white noise. Therefore, ARIMA(1,1,1) was identified as the most suitable model for forecasting pearl millet production in Jammu and Kashmir.

As shown in figure 5, the residuals of the ARIMA(1,1,1) model were randomly scat-

Table 4: ARIMA models for pearl millet production under J&K during the study period

	Training		**Validation**		**Testing**	
Models	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
ARIMA(0,1,1)	1.82	1.43	1.38	1.07	4.15	3.83
ARIMA(1,1,0)	1.81	1.42	1.18	0.97	3.91	3.57
ARIMA(1,1,1)	1.8	1.41	1.15	0.91	2.96	2.56
ARIMA(1,1,2)	1.81	1.49	1.19	0.91	3.93	3.56
ARIMA(2,1,2)	1.5	1.14	2.11	1.89	4.98	4.58
ARIMA(2,1,1)	1.75	1.42	1.2	0.98	4	3.64

Table 5: Estimates of parameters of ARIMA (1, 1, 1) for production

Model	Ljung-Box test		Parameters	Estimates	SE	p-value
	Statistics	p-value	**AR1**	0.12	0.61	0.00**
ARIMA(1,1,1)	6.86	0.73	**MA1**	-0.46	0.56	0.01**

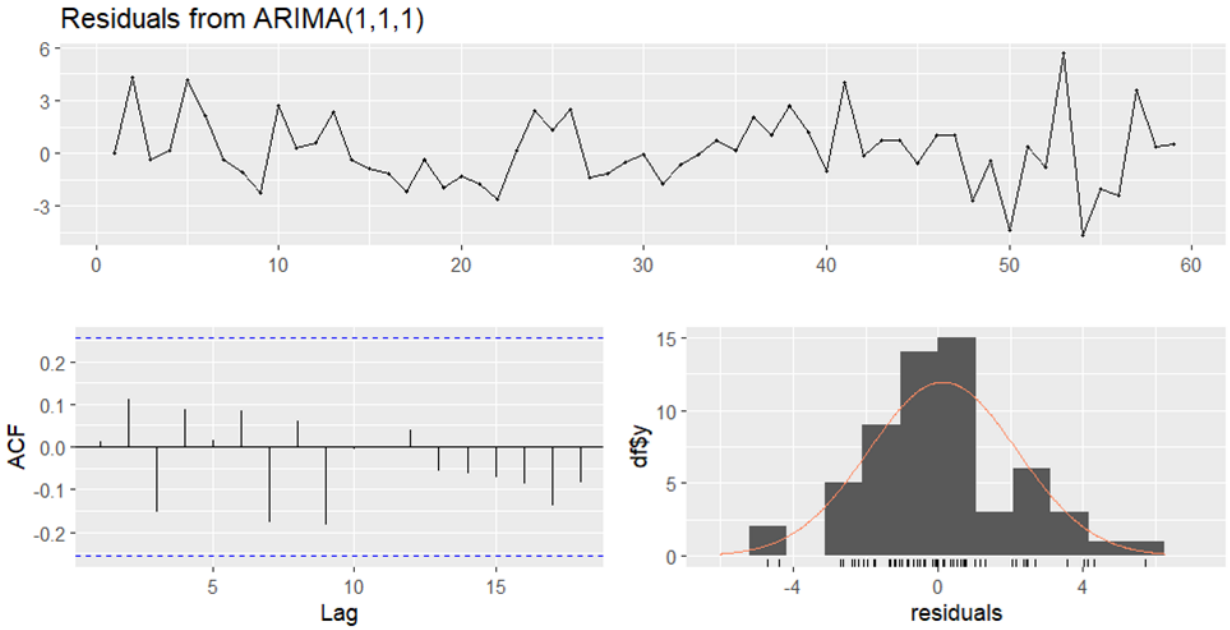


Figure 5: Residual time series plot of ARIMA (1, 1, 1) for production in J&K

tered around zero with no visible trend or systematic pattern. Although the ARIMA model adequately captured the linear pattern in the data, statistical models are developed under certain assumptions, particularly that the residuals should not contain any systematic information. Therefore, the nonlinearity of the residuals needs to be tested, which can be done using the Brock–Dechert–Scheinkman (BDS) test.

Table 6: Results of the BDS test for testing non-linearity

	Epsilon	Embedding dimension=2		Embedding dimension=3	
		Statistics	p-value	Statistics	p-value
BDS Test	0.89	2.44	0.01	-0.08	0.93
	1.78	-1.22	0.22	-2.22	0.02
	2.68	-2.63	0	-2.74	0

The BDS test was applied to the residuals of the fitted ARIMA model to examine the presence of nonlinear dependence in the pearl millet production series. As presented in Table 6, the BDS test revealed significant statistics ($p < 0.05$) for most epsilon values under embedding dimensions 2 and 3, indicating the presence of nonlinear dependence in the ARIMA residuals. These results suggest that the ARIMA model alone could not completely capture the underlying data structure, thereby justifying the use of machine learning models to capture the remaining nonlinear patterns. Therefore, machine learning models were also used, and their performance was compared with the statistical models at a later stage. The

first machine learning model employed was Nonlinear Support Vector Regression (NLSVR). For model development, the pearl millet production series was transformed into a supervised learning dataset using five lagged observations as input variables and divided into training (70%), validation (15%), and testing (15%) datasets. The lagged dataset, spanning from 1971 to 2024, was divided chronologically into training (1971 - 2006), validation (2007-2015), and testing (2016-2024) subsets to preserve temporal order and ensure a realistic evaluation of forecasting performance on future unseen data. To avoid information leakage, the data were normalized using the minimum and maximum values of the training dataset only.

Table 7: Optimal hyper-parameter values of the fitted NLSVR models

Specifications	NLSVR	ARIMA(1,1,1) residual series
No. of input lags	5	5
C	100	1
γ	0.05	0.01
ϵ	0.001	0.001

Before fitting the NLSVR model, the optimal hyperparameters were identified through a grid search procedure based on minimum validation RMSE. This NLSVR model was developed using five lagged observations, and a Radial Basis Function (RBF) kernel was used, as shown in Table 7.The optimal hyperparameters (C = 100, $\gamma = 0.05, \epsilon = 0.001$) were selected through grid search based on the minimum validation RMSE, enabling the model to effectively capture nonlinear patterns in pearl millet production data. Similarly, the Hybrid ARIMA(1,1,1)-SVR model, the residuals obtained from the fitted ARIMA(1,1,1) model were first transformed into a supervised learning dataset by considering five lagged residual observations as input variables. The dataset was subsequently divided into training (70%), validation (15%), and testing (15%) subsets, and normalization was performed using the training dataset to avoid information leakage. Before fitting the Hybrid model, the optimal hyperparameters were identified through a grid search procedure based on minimum validation RMSE. The optimal parameter combination (C = 1, $\gamma = 0.01, \epsilon = 0.001$) was identified through grid search, allowing the model to capture residual nonlinearities and enhance forecasting accuracy. Similar findings were reported by Rathod *et al.* (2018), who developed an NLSVR model for oilseed production using an RBF kernel and determined the optimal values of C = 8.19, $\gamma = 3.06$, and $\epsilon = 0.15$ through cross-validation. Their study demonstrated that systematic hyperparameter optimization considerably improved forecasting accuracy. Likewise, Pandit *et al.* (2023) reported that tuning the SVR hyperparameters before modelling the ARIMA residuals enhanced the predictive capability of hybrid forecasting models by effectively capturing nonlinear components that remained unexplained by the linear ARIMA model. Thus, the findings of the present investigation further confirm that careful hyperparameter optimization is an essential step for obtaining reliable forecasts using NLSVR and hybrid ARIMA-SVR models.

After developing the statistical and machine learning models, their forecasting performance was first compared using the Diebold–Mariano (DM) test based on the actual and predicted values. As presented in Table 8, the Diebold–Mariano test indicates that there is no statistically significant difference in forecasting accuracy among ARIMA, NLSVR, and Hybrid ARIMA–SVR models (all p-values > 0.05). Since the DM test did not identify a statistically superior model, the final model selection was based on forecast accuracy measures.

Table 8: Diebold–Mariano test results based on testing actual and predicted values

Model	DM Statistics	p-value	Accuracy Differences Between Models
Hybrid vs ARIMA	-0.5	0.31	Not Significant
Hybrid vs NLSVR	-0.73	0.23	Not Significant
NLSVR vs ARIMA	0.52	0.69	Not Significant

Table 9: Comparison of different time series models on the basis of RMSE and MAE

	Training			Validation			Testing		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
SES	1.85	1.47	18.38	1	0.84	8.55	3.01	2.6	39.76
DES	6.37	4.86	64.9	1.37	1.29	12.83	4.94	4.62	70.8
ARIMA(1,1,1)	1.8	1.41	17.96	1.15	0.91	9.65	2.96	2.56	38.94
NLSVR	0.09	0.03	4.07	0.19	1.37	11.41	0.42	0.31	36.22
ARIMA(1,1,1)-SVR	1.56	1.22	15.98	1.13	0.9	9.48	2.89	2.49	37.52

Accordingly, the performance of the SES, DES, ARIMA(1,1,1), NLSVR, and Hybrid ARIMA(1,1,1)–SVR models were evaluated using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) across the training, validation, and testing datasets, as presented in Table 9. Among the models evaluated, the NLSVR model recorded the lowest errors across training, validation, and testing datasets, indicating the highest forecasting accuracy. The Hybrid ARIMA (1,1,1)-SVR model improved upon ARIMA, while DES showed the poorest performance. Therefore, based on the minimum RMSE and MAE values obtained on the independent testing dataset, the NLSVR model was identified as the most appropriate model for forecasting pearl millet production.

Table 10: Forecasted pearl millet production using the NLSVR and SES model

Year	NLSVR	Lower Confidence Limit (95%)	Upper Confidence Limit (95%)
2026 – 2027	9.27	7.57	10.96
2027 – 2028	9.95	8.25	11.65
2028 – 2029	9.49	7.79	11.19
2029 – 2030	9.19	7.49	10.89
2030 – 2031	9.28	7.58	10.98

The NLSVR model predicts slight fluctuations in pearl millet production during 2026–2030, with forecasted values ranging from 9.19 to 9.95 thousand tonnes. Overall, production is expected to remain relatively stable, showing no significant long-term increasing or decreasing trend as presented in Table 10. Moreover, it is observed that the forecasted values through the proposed model, *i.e.*, NLSVR, with respect to the future are varying. This variation is also observed through the periodogram of the spectral analysis, as well as the ACF spikes, which show a positive trend for the initial lags and followed by negative values for several lags as shown in Figure 4.

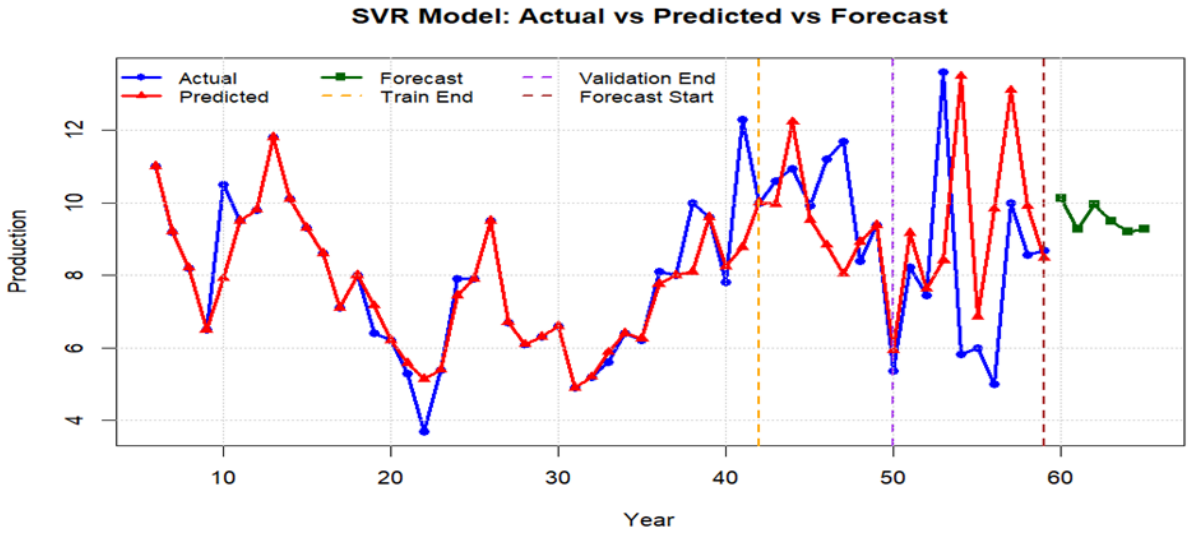


Figure 6: Graphical representation of forecasted value of pearl millet production through NLSVR

As illustrated in figure 6, the NLSVR model showed satisfactory predictive performance, with predicted values closely aligning with observed production during the training and validation phases. Despite minor deviations in the testing period, the model effectively captured the overall temporal trend and indicated relatively stable future production.

Table 11: Actual and predicted pearl millet production in the testing dataset and forecast for 2025

Year	Actual	NLSVR	Hybrid ARIMA-SVR	ARIMA	SES	DES
2015	5.36	5.94	9.07	9.75	9.59	11.59
2016	8.21	9.16	9.41	7.8	9.59	11.71
2017	7.44	7.62	9.81	8.25	9.59	11.84
2018	13.59	8.4	9.28	7.88	9.59	11.97
2019	5.82	13.48	9.43	10.51	9.59	12.1
2020	6	6.85	9.34	8.06	9.59	12.23
2021	5	9.82	9.32	7.36	9.59	12.36
2022	10	13.09	9.41	6.44	9.59	12.48
2023	8.56	9.89	9.37	8.2	9.59	12.61
2024	8.69	8.46	9.39	8.17	9.59	12.74
2025	-	10.13	9.38	8.36	8.49	5.98
χ^2		Statistics	df	Critical value (0.05, 53 df)		p-value
(value for NLSVR)		16.92	53	70.99		0.99

Table 11 presents the actual and predicted pearl millet production values in the testing dataset obtained using the different models, along with the forecast for 2025. A comparison of the predicted values with the observed production indicated that the NLSVR model exhibited the least deviation from the actual production values in the testing dataset, particularly for the most recent observation (2024), showed better predictive performance. Since the NLSVR

model demonstrated the highest forecasting accuracy based on the lowest RMSE, MAE and MAPE values on the testing dataset, it was identified as the best-performing model. To further assess the adequacy of the best-performing model, a Chi-square goodness-of-fit test was conducted for the NLSVR model. The null hypothesis (H_0) stated that there was no significant difference between the observed and NLSVR-predicted pearl millet production values, whereas the alternative hypothesis (H_1) stated that a significant difference existed between them. The test yielded a χ^2 statistic of 16.92 with 53 degrees of freedom. Since the calculated χ^2 value was substantially lower than the critical value (70.99) at the 5 per cent level of significance, and the corresponding p-value (0.99) exceeded 0.05, the null hypothesis was not rejected. These results indicate that there was no statistically significant difference between the observed and NLSVR-predicted production values. Therefore, the NLSVR model provided an adequate fit to the historical pearl millet production data and may be considered a reliable model for forecasting future production.

4. Conclusion

The present study compared the forecasting performance of statistical, machine learning, and hybrid time series models for pearl millet production in the Jammu division. The forecasting models were compared based on their predictive performance using forecast accuracy measures, including RMSE and MAE. Among the evaluated models, the NLSVR model achieved the highest forecasting accuracy and demonstrated an adequate fit to the historical data. The model forecasts indicate that pearl millet production is likely to remain relatively stable during 2026–2030, ranging from 9.19 to 9.95 thousand tonnes. Therefore, the NLSVR model can be considered a reliable tool for production forecasting and can support informed agricultural planning and policy decisions.

Acknowledgements

The authors are grateful to the Reviewer for the valuable comments on the original version of this manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Box, G. E. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*.

Department of Agriculture & Farmers Welfare. (2023). *Statewise Millet Production*. Ministry of Agriculture & Farmers Welfare, Government of India.

Directorate of Economics and Statistics, Jammu and Kashmir. (2025). *Digest of Statistics*. https://ecostatjk.nic.in/

Directorate of Millets Development. (2024). *Indian Millets Production Statistics*. Ministry of Agriculture and Farmers Welfare, Government of India. https://www.milletstats.com/

Government of India. *Economic Survey 2024–25*. New Delhi: Ministry of Finance, Department of Economic Affairs; 2025.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages (onr memorandum no. 52). *International Journal of Forecasting*, **20**, 5–10.

Holt, C. C. (1957). *Forecasting Trends and Seasonals by Exponentially Weighted Averages*. Carnegie Institute of Technology, Pittsburgh Office of Naval Research memorandum no. 52.

ICAR-All India Coordinated Research Project on Pearl Millet. *About Pearl Millet.* Jodhpur (Rajasthan), India.

Jasrotia, N., Sharma, M., Bhat, A., Mahajan, S., and Gupta, S. (2024). Forecasting models for the production of walnut in Jammu and Kashmir: A comparative study. In *Special Proceedings of the 26th Annual Conference of the Society of Statistics, Computer and Applications* (pp. 51–60). Society of Statistics, Computer and Applications. ISBN 978-81-950383-5-0.

Mahajan, S., Sharma, M., and Gupta, A. (2020). ARIMA modelling for forecasting of rice production: A case study of India. *Agricultural Science Digest-A Research Journal*, **40**, 404–407.

Mehmood, T. (2025). Agriculture as the backbone of India's economy: An analysis of contribution, barriers and policy reforms. *International Journal of Finance, Management and Economics*, **8**, 1134–1140. https://doi.org/10.33545/26179210.2025.v8.i2.660

Pal, M., Deka, J., and Rai, R. K. (1996). *Fundamentals of Cereals Crop Production*. Tata McGrew Hill Publishing Co. Ltd., New Delhi.

Pandit, P., Sagar, A., Ghose, B., Dey, P., Paul, M., Alqadhi, S., and Abdo, H. G. (2023). Hybrid time series models with exogenous variable for improved yield forecasting of major Rabi crops in India. *Scientific Reports*, **13**, 22240.

Rani, R., Khandelwal, V., Jhunjhadia, S., Ambawat, S., Bhanwariya, S., Singh, A., and Kumar, V. (2024). Quantifying genetic diversity based on morphological and molecular analysis in pearl millet (*Pennisetum glaucum [L.] R. Br.*) inbred lines. *Annals of Arid Zone*, **63**, 105–110.

Rathod, S., Singh, K. N., Patil, S. G., Naik, R. H., Ray, M., and Meena, V. S. (2018). Modeling and forecasting of oilseed production of India through artificial intelligence techniques. *Indian Journal of Agricultural Sciences*, **88**, 22-27.

Reddy, A., Yadav, O. P., and Singh, I., P. (2013). Prospects for kharif (rainy season) and summer pearl millet in Western India. Working Paper, International Crops Research Institute for Semi-Arid Tropics, Andhra Pradesh, India.

Risnawati, I., Afendi, F. M., and Sumertajaya, I. M. (2026). Performance evaluation of seasonal ARIMA-SVR and seasonal ARIMAX-SVR hybrid methods on forecasting paddy production. *BAREKENG: Journal of Mathematics and Its Applications*, **20**, 367–380.

Sharma, D. (2024). *Forecasting Models for Production of Pearl Millet (Bajra) and its Scenario in Jammu Region*. [Unpublished M.Sc. Thesis, SKUAST JAMMU].

Sharma, M., Jasrotia, N., Kumar, B., Bhat, A., and Mahajan, S. (2018). Modeling of monthly arrival of rohu fish using ARIMA in Jammu region of J&K state. *Journal of Animal Research*, **8**, 259–262.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 129-141

Estimating the Hazard Functions of Two Alternating Recurrent Events

Sugata Sen Roy[1] **and Moumita Chatterjee**[2]

[1]*Department of Statistics, St. Xavier's University Kolkata, Kolkata 700160, India*
[2]*SQC and OR Unit, Indian Statistical Institute Hyderabad, Hyderabad 500007, India*

Received: 24 July 2026; Revised: 26 August 2026; Accepted: 31 August 2026

Abstract

There has been a long history in the study of recurrent time to event data. However, very often the event itself is not instantaneous, but has a time span of its own. In such cases, the data can be looked upon as an alternating sequence of two separate recurrent events. These event times are usually correlated both over recurrences as also over the two distinct events. In this paper we review two ways of modeling such event time data.

Key words: Survival times; Alternating recurrent event; Bivariate copula; Cox's model.

AMS Subject Classifications: 62N01

1. Introduction

In life-time data analysis, one often comes across the problem of recurrent events, that is events which repeat after a certain interval of time. Examples include, relapses of the disease in individuals suffering from cancer, times to first/post-repair failures of any machine and days of periodical unemployment among agricultural labourers. The time intervals between such repetitions are usually unequal and may often exhibit certain patterns. Comprehensive reviews of early studies in this area can be found in Kalbfleisch and Prentice (2002) and Clayton (1994).

However, very often the recurrent events are not instantaneous but have themselves a life-time of their own. This gives rise to a sequence of alternating life-times. For example, when one is looking at the time to relapse of cancer, one also has intervening times of treatment, *i.e.* time to cure, before one can again start to measure the time to the next relapse. Similarly, the recurrent time-lengths of unemployment are sandwiched between recurrent times of employment, and recurrent times to machine failure are separated by recurrent repair times. Thus one would be faced with alternating sets of times for two alternating complementary events. Data of these kind are termed as "recurrent episode" data or alternating state data (Hougaard, 2000).

Corresponding Author: Sugata Sen Roy
Email: sugata.senroy@sxuk.edu.in

There have been very few studies on such alternating recurrent events. Gap time methods have been applied by Lin, Sun and Ying (1999), Wang and Chang (1999) and Pena, Strawderman and Hollander (2002), while Yan and Fine (2008) used temporal process regression to analyse recurrent episode data in cystic fibrosis patients. Sen Roy and Chatterjee (2015) analysed data of this kind without considering the correlation between the alternating events.

In this paper we compare two techniques as developed by Chatterjee and Sen Roy (2018(a) and (b)) for modelling alternating sets of times under right-censoring, where the censoring occurs primarily because of either loss to follow-up or withdrawal or the termination of the study. The first, a parametric method, uses a bivariate distribution to take account of the dependence between the two event times. The joint distribution is obtained from the marginals through a bivariate copula. Since the moments of the bivariate distribution are likely to exhibit some pattern, the parameters of the model are linked through functional relationships to capture this dependence over the cycles. In the second method, the semiparametric Cox Proportional Hazard (PH) model is used, with the dependence over the cycles and across the two events structured through an autoregressive type model.

The paper is organized as follows. Section 2 introduces the model and discusses the underlying censoring mechanism. Section 3 discusses the first method, including the choice of an appropriate copula function. The second method is discussed in Section 4. A motivation for the study was the clinical trial data on patients suffering from cystic fibrosis as reported by Fuchs *et al.* (1994). Section 5 contains a comparative study of this data based on the two methods. Some concluding remarks are made in Section 6.

2. The model

Consider two events E and F, where F is the complement of E. We also use E and F to denote the two alternating states *i.e.* an individual in state E is faced with the event F, after which he is in state F and facing event E. Assume for the moment that all individuals start in state E. We then have an alternating sequence of E's and F's. Let X and Y denote respectively the lengths of times in states E and F *i.e.* X is the time to the event F and Y is the time to the event E.

Suppose there are n individuals in the study. If for the i^{th} individual, $i = 1, \ldots, n$, there are m_i recurrences of each event, we have for this individual a sequence $X_{1i}, \cdots, X_{m_i i}$ of E-times each culminating in event F and a sequence $Y_{1i}, \cdots, Y_{m_i i}$ of F-times each culminating in event E, assuming of course that there is no censoring. (X_{ji}, Y_{ji}) then gives the complete cycle for the j^{th} recurrence with time length $X_{ji} + Y_{ji}$.

Before constructing the likelihood function some assumptions made above need to be relaxed. Firstly, it is not necessary that all individuals are in state E when they enter the study. Similarly, an individual need not necessarily be in state F with an event E occurring at the termination of his observation. In fact the observations need not end with any event and an individual may voluntarily withdraw from the study or may be lost to follow-up while in any of the two states. For example, in a study on time to cure and time to relapse, the death of an individual during the study period may be taken as a loss to follow-up. All these can be looked upon as censored observations. To incorporate these into the model we define

for the i^{th} individual the following indicator functions :

$$\delta^*_{1ji} = \begin{cases} 1 & \text{for all } \ j = 2, \cdots, m_i \\ 1 & \text{if } \ j = 1 \ \text{ and the } i^{th} \text{ individual starts in state E} \\ 0 & \text{if } \ j = 1 \ \text{ and the } i^{th} \text{ individual starts in state F} \end{cases}$$

$$\delta^*_{2ji} = \begin{cases} 1 & \text{for all } \ j = 1, \cdots, (m_i - 1) \\ 1 & \text{if } \ j = m_i \ \text{ and the } i^{th} \text{ individual withdraws in state F} \\ & \qquad \text{or ends by experiencing event E} \\ 0 & \text{if } \ j = m_i \ \text{ and the } i^{th} \text{ individual withdraws in state E} \\ & \qquad \text{or ends by experiencing event F} \end{cases}$$

While δ^*_{1ji} is an indicator of whether the i^{th} individual enters the study in the E or the F state, δ^*_{2ji} indicates whether the individual leaves the study either during or at the end of the E or F state. Further to mark whether the event is observed or censored in the last stage, we use δ_{1ji} and δ_{2ji} to indicate whether the i^{th} individual exits with an E-event or an F-event or is censored while in state E or in state F respectively, where

$$\delta_{1ji} = \begin{cases} 1 & \text{for all } \ j = 1, \cdots, (m_i - 1) \\ 1 & \text{if } \ j = m_i \ \text{ and the } i^{th} \text{ individual experiences event F} \\ 0 & \text{if the } i^{th} \text{ individual withdraws while in state E at} j = m_i \end{cases}$$

$$\delta_{2ji} = \begin{cases} 1 & \text{for all } \ j = 1, \cdots, (m_i - 1) \\ 1 & \text{if } \ j = m_i \ \text{ and the } i^{th} \text{ individual experiences event E} \\ 0 & \text{if the } i^{th} \text{ individual withdraws while in state F at} j = m_i \end{cases}$$

3. Method 1

In modelling these recurrent event-times, it is obvious that for a given individual, the X_{ji}'s and Y_{ji}'s will be dependent, both over X and Y as also over the respective sequences. However, joint modelling of (X_{ji}, Y_{ji}) for $j = 1, \ldots, m_i$ would be difficult, if not impossible. An alternative would be to begin by assuming that the sequence of X_{ji}'s and Y_{ji}'s have the same distributional form, say, $F_X(x, \theta_j)$ and $F_Y(y, \eta_j)$ respectively, where $\theta_j = (\theta_{1j}, \ldots, \theta_{pj})'$ and $\eta_j = (\eta_{1j}, \ldots, \eta_{qj})'$ are respectively the p and q dimensional parameter vectors which characterize the two distribution functions. To accommodate the relationships between the X_{ji}'s, let, for each $k = 1, \ldots, p$, $g_k(h, \alpha_k)$ be continuous and differentiable functions of h which are characterized by the r_k-dimensional parameter vectors α_k's. Then for $j = 1, \ldots, m_i$, the parameters $\theta_{\mathbf{j}}$ are linked through $\theta_{kj} = g_k(j, \alpha_k)$, $k = 1, \ldots, p$. Similarly, the relationships between the Y_{ji}'s are modelled through the parameters as $\eta_{lj} = h_l(j, \beta_l)$, $l = 1, \ldots, q$ and $j = 1, \ldots, m_i$, with β_l's the s_l-dimensional parameter vectors characterizing the continuous and differentiable functions h_l for $l = 1, \ldots, q$.

The g_k and h_l functions are not predetermined. For a given data-set, estimates of the parameters θ_j are first obtained by separately fitting $F_X(x, \theta_j)$ for each of the cycles, $j = 1, \ldots, m_i$. Then plotting each estimated θ_{kj} against j, an idea of the functional form of $g_k(j, \alpha_k)$ for $k = 1, \ldots, p$ can be envisaged. The functional forms of the elements of h_l can be obtained similarly.

The g_k and h_l functions allow us to capture the pattern that the two events exhibit over the cycles. However, the inter-relationship between the X_{ji} and the Y_{ji} are still to be accounted for, and this we do through the use of a copula.

As is well-known, a copula is a function through which a multivariate distribution function can be built from the univariate marginal distributions of each of the variables. A bivariate copula can thus be used in our model to find a functional relationship between X_{ji} and Y_{ji} for each j. Thus if C denotes the copula function, then under the assumption of continuity of F_{X_j} and F_{Y_j}, it follows from Sklar (1959) that the joint distribution of X_{ji} and Y_{ji} is

$$F_{X_jY_j}(x, y, \zeta_j) = C(F_{X_j}(x, \theta_j), F_{Y_j}(y, \eta_j), \xi_j), \tag{1}$$

where ξ_j is a dependence parameter and $\zeta_j = (\theta_j{}', \eta_j{}', \xi_j)'$.

Let $C'_U(u, v)$, $C'_V u, v)$ and $C''_{UV}(u, v)$ denote respectively the first and second derivatives of C with respect to u and v. Defining the marginal probability functions of X_{ji} and Y_{ji} respectively as $f_{X_j}(x, \theta_j)$ and $f_{Y_j}(y, \eta_j)$, the joint probability density function of (X_{ji}, Y_{ji}) is then

$$f_{X_jY_j}(x_{ji}, y_{ji}, \zeta_j) = C''_{F_X F_Y}(F_{X_j}(x_{ji}, \theta_j), F_{Y_j}(y_{ji}, \eta_j), \xi_j) f_{X_j}(x_{ji}, \theta_j) f_{Y_j}(y_{ji}, \eta_j). \tag{2}$$

Since by our assumption, X_{ji} precedes Y_{ji}, the survival and hazard functions of X_{ji} can be obtained directly from the marginal distribution function as

$$S_{X_j}(x_{ji}, \theta_j) = 1 - F_{X_j}(x_{ji}, \theta_j) \tag{3}$$

$$\text{and} \quad \lambda_{X_j}(x_{ji}, \theta_j) = -\frac{d}{dx_{ji}} ln S_{X_j}(x_{ji}, \theta_j), \tag{4}$$

respectively. However, the survival and hazard functions for Y_{ji} will be conditioned on $X_{ji} = x_{ji}$. Since the conditional distribution of $Y_{ji}|X_{ji} = x_{ji}$ is

$$F_{Y_j|X_j}(y_{ji}, \zeta_j|x_{ji}) = C'_{F_X}(F_{X_j}(x_{ji}, \theta_j), F_{Y_j}(y_{ji}, \eta_j), \xi_j) \tag{5}$$

these are given respectively by

$$S_{Y_j|X_j}(y_{ji}, \zeta_j|x_{ji}) = 1 - F_{Y_j|X_j}(y_{ji}, \zeta_j|x_{ji}) \tag{6}$$

$$\text{and} \quad \lambda_{Y_j|X_j}(y_{ji}, \zeta_j|x_{ji}) = -\frac{d}{dy_{ji}} ln S_{Y_j|X_j}(y_{ji}, \zeta_j|x_{ji}). \tag{7}$$

However, like the θ_{ji}'s and η_{ji}'s, the dependence parameters ξ_j's are also likely to be related over the cycles for $j = 1, \ldots, m_i$. This is accounted for through the relation $\xi_j = u(j, \kappa)$, where $\kappa = (\kappa_1, \ldots, \kappa_d)'$ is the d-dimensional parameter vector characterizing the continuous and differentiable function u.

We define the $r = \sum_{k=1}^{p} r_k$ dimensional vector $\alpha = (\alpha'_1, \ldots, \alpha'_p)'$ and the $s = \sum_{l=1}^{q} s_l$ dimensional vector $\beta = (\beta'_1, \ldots, \beta'_q)'$ and write $\nu = (\alpha', \beta', \kappa)'$ as the $(r + s + d)$-dimensional vector of unknown parameters. The likelihood function is then constructed by considering for each individual and each cycle the marginal survival and hazard functions of the E-times

X_{ji} and the conditional survival and hazard times of the F-times $Y_{ji}|X_{ji} = x_{ji}$. This leads to

$$\begin{aligned} l(\nu) &= l(\alpha, \beta, \kappa|(x_{ji}, y_{ji}), j = 1 \ldots, m_i, i = 1, \ldots n) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m_i} \left[\left\{ S_{X_j}(x_{ji}, \mathbf{g}(j,\alpha)) \left\{ \lambda_{X_j}(x_{ji}, \mathbf{g}(j,\alpha)) \right\}^{\delta_{1ji}} \right\}^{\delta^*_{1ji}} \right. \\ &\left. \left\{ S_{Y_j|X_j}(y_{ji}, \mathbf{g}(j,\alpha), \mathbf{h}(j,\beta), u(j,\kappa)) \left\{ \lambda_{Y_j|X_j}(y_{ji}, \mathbf{g}(j,\alpha), \mathbf{h}(j,\beta), u(j,\kappa)) \right\}^{\delta_{2ji}} \right\}^{\delta^*_{2ji}} \right] \end{aligned} \tag{8}$$

Building the score vector and the information matrix by taking first and second derivatives with respect to $\nu = (\alpha', \beta', \kappa)'$, the maximum likelihood (m.l.) estimates $\hat{\nu} = (\hat{\alpha}', \hat{\beta}', \hat{\kappa})'$ can now be obtained simultaneously by the Newton-Raphson method. This for any cycle j leads to the estimator $\hat{\zeta}_j = (\hat{\theta}'_j, \hat{\eta}'_j, \hat{\xi}_j)'$ of the model parameters and hence to the estimators $\hat{S}_{X_j}(x_{ji}, \hat{\theta}_j)$ and $\hat{S}_{Y_j|X_j}(y_{ji}, \hat{\zeta}_j|x_{ji})$ of the respective survival functions and their corresponding hazard functions.

The likelihood function (8) depends on the copula through equations (3-7). Hence it is necessary to choose an appropriate copula before maximizing the likelihood function. Owing to its simple form which is easy to work with, the Archimedean copula is a popular choice for modelling bivariate distributions (Nelson, 2006). Its added advantage is that it caters to a variety of dependence structures including Kendall's τ, which can be evaluated directly from its generator Genest and Mackay (1986). Shih (1998) remarks that in survival studies, the Clayton copula, a member of the Archimedean family, is probably the most widely used, primarily because of its mathematical convenience and attractive properties. In this paper we tried different copulas under the Archimedean set-up, but eventually found that the Clayton copula gives the best fit.

4. Method 2

Next let the E-times depend on a set of p covariates $\mathbf{z} = (z_1, \ldots, z_p)$ while the F-times depend on a set of q covariates $\mathbf{w} = (w_1, \ldots, w_q)$. Then the Cox PH models for the hazard functions for the j^{th} cycle of the E and F times are

$$\begin{aligned} \lambda_{Ej}(x_{ji}, \boldsymbol{z}_j) &= \lambda_{E0j}(x_{ji}) e^{\boldsymbol{\theta}'_j \boldsymbol{z}_j} \\ \lambda_{Fj}(y_{ji}, \boldsymbol{w}_j) &= \lambda_{F0j}(y_{ji}) e^{\boldsymbol{\eta}'_j \boldsymbol{w}_j} \end{aligned}$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \cdots, \theta_{jp})'$ and $\boldsymbol{\eta}_j = (\eta_{j1}, \cdots, \eta_{jq})'$ are the p and q dimensional vectors of coefficients, and $\lambda_{E0j}(x)$ and $\lambda_{F0j}(y)$ are the baseline hazard functions. Of course, in our study, there are no exogenous covariates. However, to account for the dependence over the alternating sequences, we assume a first order autoregressive-type model where x_j depends on y_{j-1} and x_{j-1}, while y_j depends on x_j and y_{j-1}. Of course, this can be easily extended to a higher-order model, but for simplicity and because the number of cycles will generally be small, we restrict ourselves to the first order only. Defining $x_0 = 0$ and $y_0 = 0$, this leads

to the hazard model

$$\begin{aligned}
\lambda_{Ej}(x_{ji}, \boldsymbol{z}_{ji}, \mathbf{y}_{ji}) &= \lambda_{E0j}(x_{ji}) e^{(\mu_1 y_{(j-1)i} + \mu_2 x_{(j-1)i})} \\
&= \lambda_{E0j}(x_{ji}) e^{\boldsymbol{\mu}' \mathbf{y}_{ji}} \\
\lambda_{Fj}(y_{ji}, \boldsymbol{w}_{ji}, \mathbf{x}_{ji}) &= \lambda_{F0j}(y_{ji}) e^{(\gamma_1 x_{ji} + \gamma_2 y_{(j-1)i})} \\
&= \lambda_{F0j}(y_{ji}) e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}}
\end{aligned}$$

where $\mathbf{y}_{ji} = (y_{(j-1)i}, x_{(j-1)i})'$, $\mathbf{x}_{ji} = (x_{ji}, y_{(j-1)i})'$, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$.

The corresponding survival functions are

$$\begin{aligned}
S_{Ej}(x_{ji}, \boldsymbol{z}_{ji}, \mathbf{y}_{ji}) &= [S_{E0j}(x_{ji})]^{e^{\boldsymbol{\mu}' \mathbf{y}_{ji}}} \\
S_{Fj}(y_{ji}, \boldsymbol{w}_{ji}, \mathbf{x}_{ji}) &= [S_{F0j}(y_{ji})]^{e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}}}
\end{aligned}$$

where $S_{E0j}(x_{ji})$ and$S_{F0j}(y_{ji})$ are the respective baseline survival functions.

For $m = \max m_i$, the likelihood function based on the m_i recurrences of each of the n individuals is then

$$\begin{aligned}
& l(\boldsymbol{\mu}, \boldsymbol{\gamma} | (x_{ji}, y_{ji}), j = 1 \ldots, m_i, i = 1, \ldots n) \\
= & \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\left\{ S_{Eji}(x_{ji}, \mathbf{y}_{ji}) \left\{ \lambda_{Eji}(x_{ji}, \mathbf{y}_{ji}) \right\}^{\delta_{1ji}} \right\}^{\delta^*_{1ji}} \left\{ S_{Fji}(y_{ji}, \mathbf{x}_{ji}) \left\{ \lambda_{Fji}(y_{ji}, \mathbf{x}_{ji}) \right\}^{\delta_{2ji}} \right\}^{\delta^*_{2ji}} \right] \\
= & \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\{ \lambda_{E0j}(x_{ji}) e^{\boldsymbol{\mu}' \mathbf{y}_{ji}} \}^{\delta_{1ji} \delta^*_{1ji}} \{ e^{-\Lambda_{E0j}(x_{ji}) e^{\boldsymbol{\mu}' \mathbf{y}_{ji}}} \}^{\delta^*_{1ji}} \right] \times \\
& \left[\{ \lambda_{F0j}(y_{ji}) e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}} \}^{\delta_{2ji} \delta^*_{2ji}} \{ e^{-\Lambda_{F0j}(y_{ji}) e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}}} \}^{\delta^*_{2ji}} \right] \\
= & \; l_E(\boldsymbol{\mu}, | (x_{ji}, y_{ji}), j = 1 \ldots, m, i = 1, \ldots n) l_F(\boldsymbol{\gamma} | (x_{ji}, y_{ji}), j = 1 \ldots, m, i = 1, \ldots n), \quad (9)
\end{aligned}$$

$\Lambda_{E0j}(x_{ji})$ and $\Lambda_{F0j}(y_{ji})$ being the baseline cumulative hazard functions of E and F, respectively.

Equation (9) shows that the likelihood function decomposes into distinct E and F components. Hence treating the two components separately, and interchanging the order of the product, the partial likelihoods for the two components are obtained as

$$pl_E(\boldsymbol{\mu}) = \prod_{j=1}^{m} \prod_{i=1}^{n_{Ej}} \left[\frac{e^{\boldsymbol{\mu}' \mathbf{y}_{ji}}}{\sum_{l \varepsilon \Re_{Ej}(x_{ji})} e^{\boldsymbol{\mu}' \mathbf{y}_{ji}}} \right]^{\delta_{1ji} \delta^*_{1ji}} \quad (10)$$

$$\text{and} \quad pl_F(\boldsymbol{\gamma}) = \prod_{j=1}^{m} \prod_{i=1}^{n_{Fj}} \left[\frac{e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}}}{\sum_{l \varepsilon \Re_{Fj}(y_{ji})} e^{\boldsymbol{\gamma}' \mathbf{x}_{ji}}} \right]^{\delta_{2ji} \delta^*_{2ji}} \quad (11)$$

where n_{Ej} and n_{Fj} are respectively the number of individuals in the risk sets $\Re_{Ej}$ and $\Re_{Fj}$ at the beginning of the E and F states of the j^{th} cycle.

In order to find the estimates of the parameters involved we treat the two equations (10) and (11) separately, and obtain their score vectors and information matrices. The Newton-Raphson method can then be used to solve for the estimates.

5. Numerical examples

5.1. The rhDNase data

In this section we analyse the cystic fibrosis data as reported by Fuchs et. al. (1994). Cystic fibrosis is a genetic disorder that can cause chronic infection. The infection produces thick, sticky mucus that clogs the respiratory organs and can recur several times in a patient's lifetime. Each infection can have different times to cure. In 1992, a trial was conducted to test the extent to which rhDNase could reduce the clogging as compared to placebo. Of the 647 patients, 322 were placed under the treatment and their times to cure as well as times to the next infection were recorded. Several recurrences are observed for some of the patients. However, data beyond the 3^{rd} recurrence is sparse. Fuchs et. al. (1994) restricted their study to the first exacerbation only.

Therneau and Hamilton (1997) considered the successive exacerbation although as independent occurrences. Chang and Wang (1999) and Huang and Wang (2005) implemented the Gap time method to analyze the episode lengths but considered them independently of the frequency. Moreover, their assumption of conditional independence of the gap times did not match the empirical findings. However, none of the previous studies considered estimating the survival times of both infection and cure assuming a dependence structure.

In using our technique we assume the *onset of the infection* as event F (state E being the non-infection period) and *cure* as the event E (state F being the period of infection).

5.2. Method 1

In the first method, to guess the underlying distributions, the Kaplan-Meier (KM) estimator (Kleinbaum and Klein, 2012) is used to obtain rough fits to the data for both the infection and cure times for each of the cycles. Four alternative distributions, the lognormal (LN), the Loglogistic (LL), the Gamma (G) and the Weibull (W) are then fitted to each of these times and these are mapped on to the corresponding KM curves. The goodness of fit for each distribution is judged using a technique common in functional data analysis. This requires recording the survival probabilities at several selected time-points on each of the five fitted curves. The sum of squares (SSE) of the difference between these values for each of the four curves against the corresponding values on the KM curve are then computed. The curve with the minimum SSE is then selected. For our data, the Weibull gives the best fit for both the time to infection and the time to cure for most of the cycles and particularly for the first three. Since the data from the fourth cycle onwards is sparse, the Weibull distribution is selected to model the two event times.

Writing the marginal distribution of the j^{th} time to infection (X_j) as $W(\theta_{1j}, \theta_{2j})$, $\theta_{1j}, \theta_{2j} > 0$ and the marginal distribution of the j^{th} time to cure (Y_j) as $W(\eta_{1j}, \eta_{2j})$, $\eta_{1j}, \eta_{2j} > 0$, individual Weibull distributions are fitted to each cycle to get initial estimates of the scale parameters θ_{1j} and η_{1j} and the shape parameters θ_{2j} and η_{2j} for each j. A plot of each of these against j is shown in Figure 1. Although the plot for η_1 is not very conclusive, we assume the following relationships:

$$\begin{aligned} \theta_{1j} &= e^{(\alpha_{11}+\alpha_{12}j)} \quad \text{and} \quad \theta_{2j} = e^{(\alpha_{21}+\alpha_{22}j)}; \\ \eta_{1j} &= e^{(\beta_{11}+\beta_{12}j)} \quad \text{and} \quad \eta_{2j} = e^{(\beta_{21}+\beta_{22}j)}; \quad j = 1, \cdots, 4. \end{aligned} \tag{12}$$

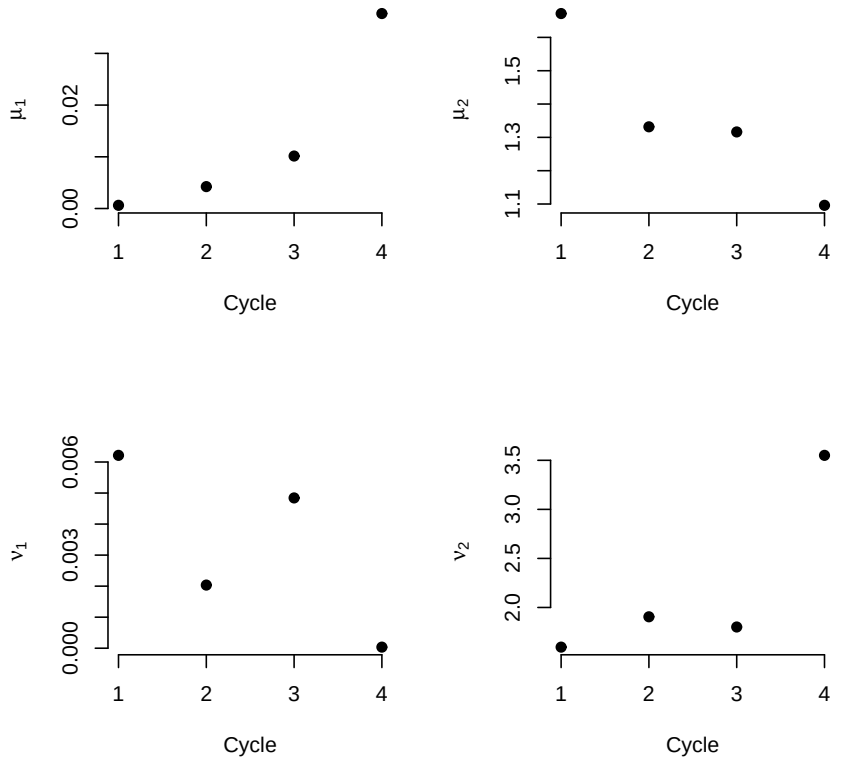


Figure 1: Functional forms of the estimated Weibull parameters over the stages

Next taking ξ_j as the dependence parameter, the joint distribution of (X_j, Y_j) is constructed for the j^{th} cycle using the Clayton copula.

Computing Kendall's $\hat{\tau}_j$ for each j, where $\hat{\tau}_j = (C_j - D_j)/(C_j + D_j)$, C_j and D_j being respectively the number of concordant and discordant pairs in the j^{th} cycle, the estimates of $\hat{\xi}_j$ are obtained from the relation $\hat{\xi}_j = 2\hat{\tau}_j/(1-\hat{\tau}_j)$, which holds true for the Clayton copula. The values across the cycles indicate that a constsnt ξ_j =-0.28 can be assumed for all j. The negative value of $\hat{\xi}_j$ is justified since it is expected that the more severe is the case, the longer is the times to cure and the shorter is the time to the next relapse.

Writing the likelihood function in (8) in terms of $(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})'$ and solving by Newton Raphson method, the estimates as shown in Table 1 are obtained. The corresponding estimated standard errors are also shown in brackets. Finally, the estimated hazard curves for infection and cure are obtained as shown in Figure 2.

The left panel relates to the time to infection, and both for placebo and treatment, the hazards are increasing over cycles, which means that more the recurrences the quicker is the chance of the next recurrence. For the placebo, although initially decreasing slightly, the hazards remain, more or less constant, over time. However, for the treatment, except for the first cycle, the hazards are intially low but increases over time. This means that with time the chances of catching an infection increases. What intrigues us is that for the

second and third cycle the hazards for treatment are greater than that of the palcebo after a given time. This is counter-intutive and needs to be more closely looked into. For time to cure, the right panel shows increasing hazards for both placebo and treatment over all the cycles, with the latter dominating the former in general. This implies that cure under treatment is faster than that under placebo. And as expected, the hazards for both placebo and treatment increases over cycles, implying that in either case, with successive recurrences, it takes longer to get cured.

Table 1: Estimates of the parameters and their standard errors along with 95% C.I.

Time to infection	**Estimates**	**std. error**	**95% C.I.**
α_{11}	-3.4152	0.6560	(-3.4868,-3.34354)
α_{12}	-0.5144	0.4103	(-0.5592,-0.4696)
α_{21}	0.294	0.1416	(0.2785,0.3095)
α_{22}	-0.2357	0.0814	(-0.2446,-0.2268)
Time to cure	**Estimates**	**std. error**	**95% C.I.**
β_{11}	2.8855	0.7307	(2.8057,2.9653)
β_{12}	2.3647	0.514	(2.3085,2.4208)
β_{21}	0.2239	0.1058	(0.2123,0.2355)
β_{22}	0.3011	0.0523	(0.2954,0.3068)

5.3. Method 2

In this section, we use the Cox model with the autoregressive-type event-time structures. Table 2 shows the autoregressive coefficients obtained by solving the partial likelihoods (10)-(11). Joint testing of $H_{01} : (\mu_1, \mu_2) = (0, 0)$ and $H_{02} : (\gamma_1, \gamma_2) = (0, 0)$ shows that the current times of relapse and cure are significantly dictated by their past values. In Figure 3, the hazard curves are plotted against the baseline hazard for both treatment and placebo.

Intersetingly, the left panel of Figure 3 corroborates the fact that hazard of infection is greater under the treatment than under the placebo. However, as in method 1, the cure-time is faster under treatment than under placebo for all the cycles.

Table 2: The estimates of the regression parameters and their standard errors

parameters	**Estimates**	**standard errors**
(μ_1, μ_2)	(0.09, 0.009)	(0.003, 0.001)
(γ_1, γ_2)	(0.02, 0.06)	(0.002, 0.004)

6. Some concluding remarks

In this paper we have considered two different methods for modelling alternating recurrent event times. The first models the dependence structure through a bivariate copula and relationships of the cycle parameters, while the second uses the Cox PH model with the dependence brought in through autoregressive structures. The methods are illustrated

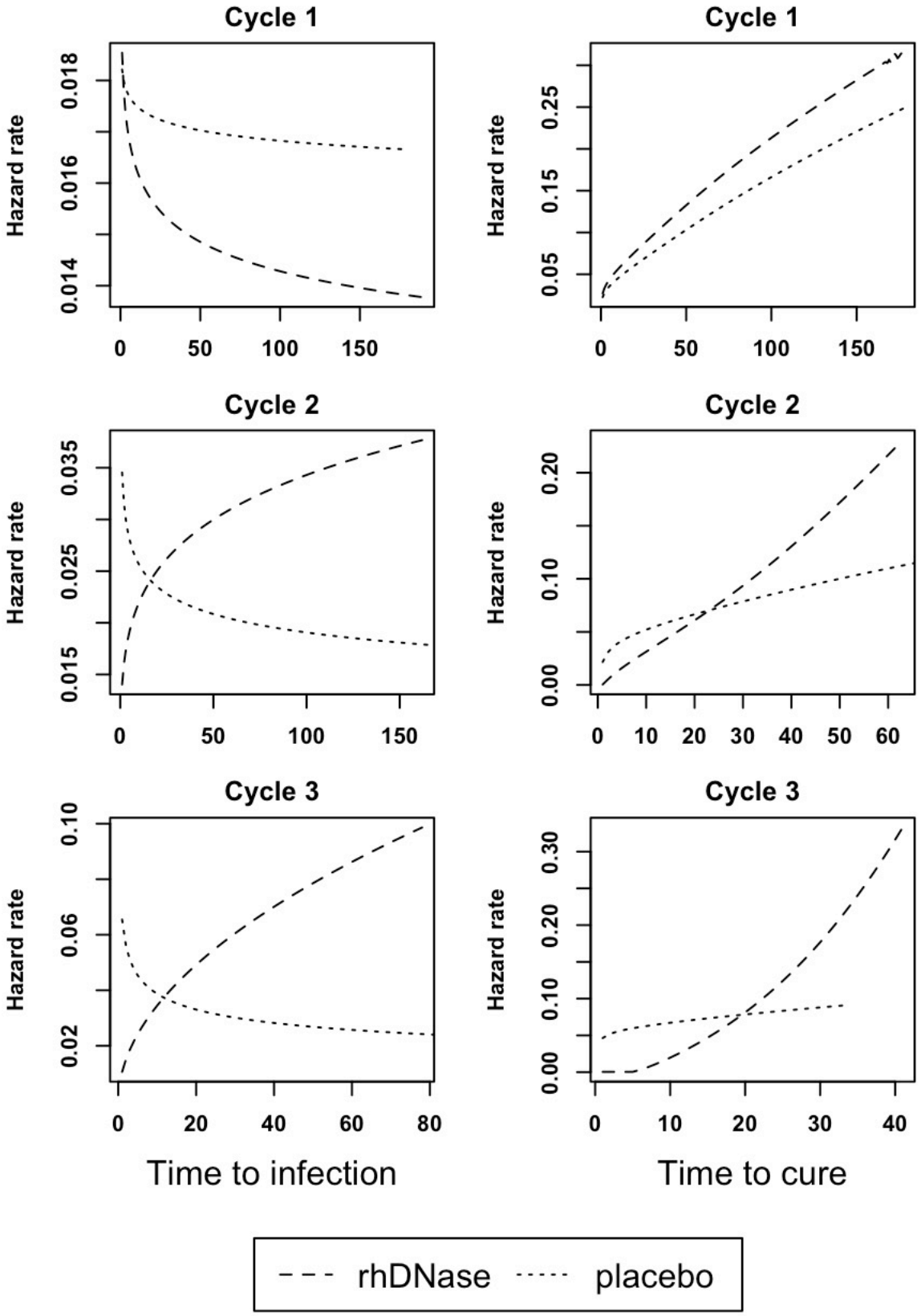


Figure 2: The hazard curves by method 1

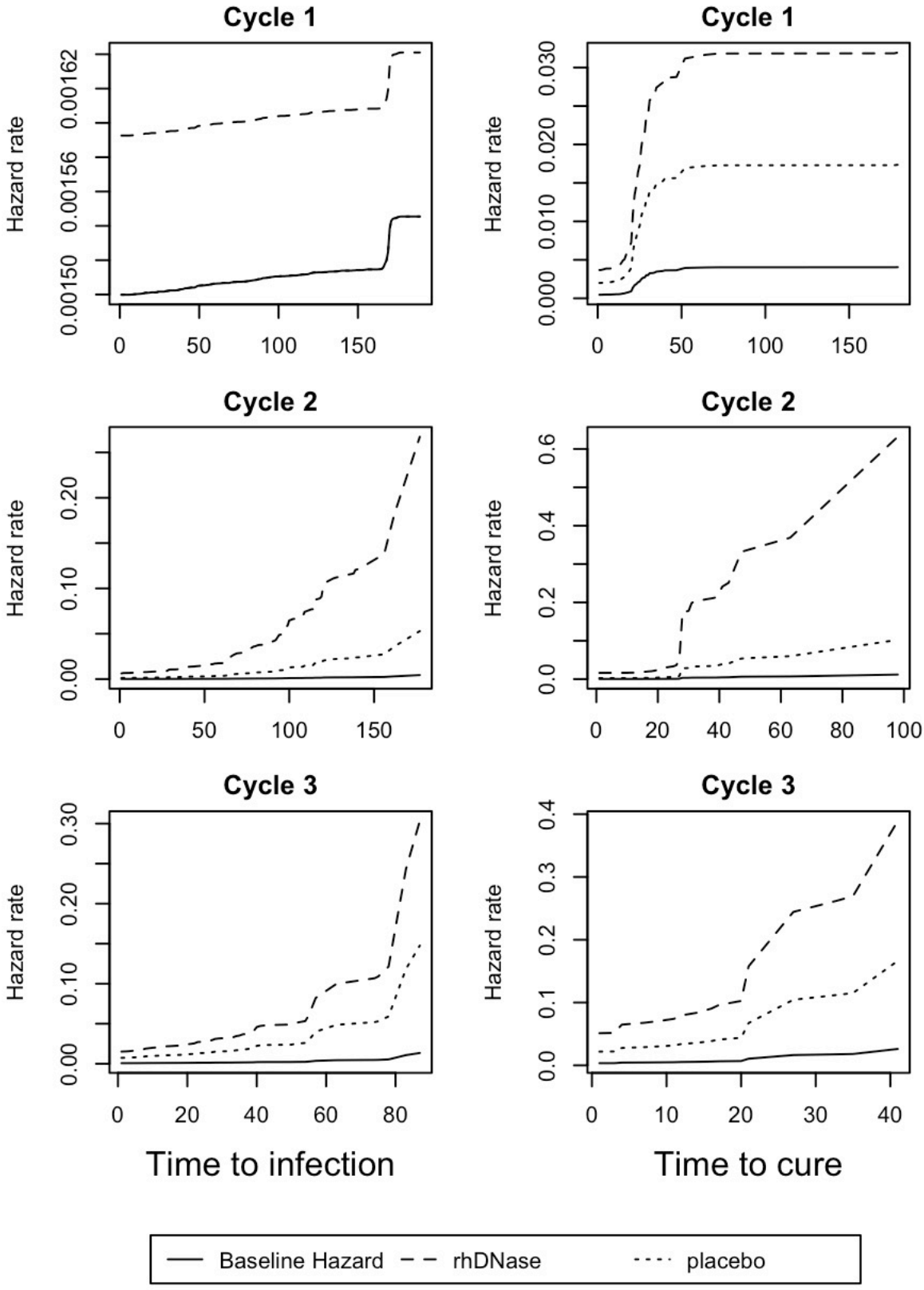


Figure 3: The hazard curves by method 2

through the cystic fibrosis data-set. Except for the first-cycle time to infection under treatment, the patterns in each of the two methods corroborate each other.

The first method, since it directly models the pattern over cycles, is likely to give smoother and sequenced hazard curves. However, if one is interested in the small twists and turns, the second method might be preferred.

The study can be extended to other forms of censoring, particularly interval censoring. Also covariates can be incorporated in the models to observe their impacts on survival times.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Chang, S. H. and Wang, M. C. (1999). Conditional regression analysis for recurrence-time data. *Journal of the American Statistical Association*, **94**, 1221–1230.

Chatterjee, M. and Sen Roy, S. (2018a). Estimating the hazard functions of two alternating recurrent events in the presence of covariates. *Advances in Statistical Analysis*, **102**, 289-304.

Chatterjee, M. and Sen Roy, S. (2018b). A copula-based approach for estimating the survival functions of two alternating recurrent events. *Journal of Statistical Computing and Simulation*, **88**, 3098-3115.

Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research*, **3**, 244–262.

Fuchs, H., Borowitz, D., Christiansen, D., Morris, E., Nash, M., Ramsey, B., Rosenstein, B., Smith, A., and Wohl, M. (1994). Effect of aerosolized re-combinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. *New England Journal of Medicine*, **331**, 637–642.

Genest, C. and MacKay, R. J. (1986). The joy of copulas: Bivariate distributions with univariate marginals. *The American Statistician*, **40**, 280–283.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. 1^{st} ed. (NY): Springer-Verlag.

Huang, C. Y. and Wang, M. C. (2005). Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics.*, **61**, 392-402.

Kalbfleisch, J. D. and Prentice, R. L.(2002). *The Statistical Analysis of Failure Time Data*. 2^{nd} ed. John Wiley & Sons.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-learning Text*. 3^{rd} ed. Springer, New York.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of gap time distributions for serial events with censored data. *Biometrika*, **86**, 59–70.

Nelson, R. B. (2006). *An Introduction to Copulas*. 2^{nd} ed. Springer, New York.

Pena, E. A., Strawderman, R. L., and Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association*, **96**, 1299–1315.

Sen Roy, S. and Chatterjee, M. (2015). Estimating the hazard functions of two alternating occurring recurrent events. *Journal of Applied Statistics*, **42**, 1547–1555.

Shih, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, **85**, 189–200.

Sklar, A. (1959), Fonctions de répartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

Therneau, T. M. and Hamilton, S. A.. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, **16**, 2029–2047.

Wang, M. C. and Chang, S. H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, **94**, 146–153.

Yan, J. and Fine, J. P. (2008). Analysis of episodic data with application to recurrent pulmonary exacerbations in cystic fibrosis patients. *Journal of the American Statistical Association*, **103**, 498–510.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 143-168

Valuing Free Digital Services in India: Experimental Consumer-Welfare Estimates and the 2025 SNA Framework*

Gulab Singh[1] and A. C. Kulshreshtha[2]

[1]*Formerly with Central Statistics Office, Government of India and United Nations Statistics Division, New York, USA*
[2]*Formerly with Central Statistics Office, Government of India, and UNSIAP, Japan*

Received: 15 August 2026; Revised: 25 August 2026; Accepted: 01 September 2026

Abstract

The rapid expansion of zero-price digital services has widened the divergence between market-based measures of economic production and important dimensions of consumer welfare, creating a measurement gap that conventional national accounts are not designed to close. This paper develops an experimental framework for assessing the consumer-welfare value of free digital services in India and examines how such estimates relate to the measurement architecture of the 2025 System of National Accounts (2025 SNA). Willingness-to-accept (WTA) valuation is used as the principal approach, supplemented by time-use valuation and an advertising-revenue calibration benchmark. In the absence of a nationally representative Indian WTA survey, service-specific valuations are transferred from international experimental evidence and adjusted to the Indian context; the resulting estimates should therefore be interpreted as indicative orders of magnitude rather than definitive welfare aggregates. The principal specification yields a median-equivalent consumer-welfare estimate of approximately ₹10.4 lakh crore in 2023–24, equivalent to about 3.4 per cent of nominal GDP, subject to substantial parameter and transfer uncertainty. These welfare estimates are conceptually distinct from national-accounting output. Under the 2025 SNA, consumer surplus remains outside core GDP, while complementary extended accounts can make household consumption of free platform services visible using producer cost as the preferred valuation basis. The paper therefore distinguishes clearly among core SNA GDP, cost-valued extended accounting and consumer-welfare measurement. It concludes by identifying priorities for strengthening India's statistical infrastructure, particularly nationally representative incentive-compatible valuation experiments, improved digital time-use data and information on platform production cost.

*This paper is based on the special remarks made by the second author, Dr. A. C. Kulshreshtha, in the Financial Statistics Section of the Pune Conference.

Key words: Consumer surplus; Free digital services; GDP measurement; National accounts; 2025 SNA, Digital economy; Welfare measurement.

Corresponding Author: A. C. Kulshreshtha
Email: ackulshreshtha@yahoo.com

1. Introduction

National income accounting was designed for an economy in which value creation occurred predominantly through priced market transactions. The System of National Accounts — codified in 1953 and revised through 1968, 1993, 2008, and 2025 — defines its production boundary primarily around goods and services exchanged at economically significant prices, with consumer surplus excluded by design rather than by oversight. This framework has served well for decades, but the digital transformation has progressively eroded the correspondence between market-measured output and economic welfare, as consumers increasingly rely on services offered at zero monetary price.

India provides a particularly important setting for examining this measurement problem. Over the past decade, the country has experienced exceptionally rapid expansion in internet access, smartphone adoption and use of digital platforms, resulting in a user base of several hundred million people. Much of this digital consumption takes place through services for which users pay no explicit monetary price, including search, messaging, social media, online video, digital maps and selected digital-payment services.

This paper makes three principal contributions. First, it develops an experimental framework for assessing the consumer welfare associated with free digital services in India. Given the absence of nationally representative Indian valuation experiments, the paper combines transferred WTA evidence with Indian user-base and usage information and explicitly quantifies the uncertainty surrounding the resulting estimates. The estimates are therefore intended to establish plausible orders of magnitude rather than definitive welfare aggregates. Second, the paper places these estimates within the conceptual architecture of the 2025 SNA and distinguishes among three fundamentally different measures: core SNA GDP, cost-valued extended accounts for free platform services, and consumer-surplus measures that lie outside the production boundary. Third, it identifies the statistical information that would be required for India to move from experimental estimates toward regular measurement, including platform production costs, service-specific usage information and nationally representative incentive-compatible WTA surveys. In this way, the paper seeks to contribute both to the empirical measurement of India's digital economy and to the broader discussion of how conventional production accounts can be complemented by welfare indicators in an increasingly digital economy.

The paper proceeds as follows. Section 2 reviews the theoretical foundations and relevant literature. Section 3 describes India's free digital-services ecosystem and the data used in the analysis. Section 4 presents the estimation framework and associated uncertainty analysis. Section 5 situates the resulting estimates within the 2025 SNA architecture, distinguishing core accounts, extended accounts and welfare indicators. Section 6 considers the implications for welfare-augmented growth measurement without constructing a numerical GDP-B series. Section 7 places the Indian estimates in an international context. Section 8 discusses policy and statistical implications, while Section 9 presents sensitivity analysis. Section 10 sets out the principal limitations and priorities for future empirical work. Section 11 concludes.

2. Theoretical foundations and literature

2.1. The measurement problem

The measurement challenge arises because national accounts rely on market prices as the primary valuation basis and define the production boundary around activities producing outputs for market exchange. When a service carries zero monetary price, it generates no consumer expenditure and contributes nothing to measured household final consumption. The platform's advertising revenue is recorded as intermediate consumption by purchasing businesses or as the platform's output of advertising services — not as household final consumption of the free service. The welfare-generating activity on the consumer side remains unrecorded.

The literature has long recognised that GDP measures production, not welfare, and that consumer surplus is excluded by design (Nordhaus, 2006; Stiglitz *et al.*, 2009). When consumption shifts systematically toward zero-price services, GDP's utility as a welfare proxy may deteriorate, though it remains a valid production measure.

2.2. Valuation approaches for zero-price services

Three broad approaches have emerged in the literature. The willingness-to-accept (WTA) approach, developed most rigorously by Brynjolfsson *et al.* (2019), elicits the minimum compensation consumers would require to forgo access to a digital service. This directly estimates Hicksian equivalent variation — the theoretically appropriate welfare measure. However, WTA estimates face hypothetical bias (stated valuations exceeding revealed valuations by a factor of 2–3; List and Gallet, 2001; Murphy *et al.*, 2005), anchoring effects, and incentive compatibility challenges.

A further, related concern is the response problem inherent in WTA elicitation: stated valuations are sensitive not only to hypothetical bias but also to survey non-response and self-selection into valuation experiments, question framing and elicitation-format effects, and strategic responses where incentive compatibility is imperfect. The correction factor β applied in Section 4.2 addresses the hypothetical-bias component specifically — the empirically documented gap between stated and incentive-compatible valuations — but it is calibrated from international meta-analyses and does not, and cannot, correct for response and framing effects specific to an Indian survey instrument that has not yet been fielded. Because these are properties of the elicitation instrument rather than of the underlying construct being measured, they cannot be resolved analytically; they require methods that are less susceptible to hypothetical bias by design. Two complementary approaches are particularly promising for India. Discrete choice experiments, which infer valuations from respondents' choices among bundled alternatives rather than from direct stated-value questions, are less prone to anchoring and framing effects than open-ended WTA elicitation. Randomised controlled trials that vary actual access to a digital service — in the spirit of Allcott and Gentzkow (2019) and the deactivation experiments underlying Brynjolfsson *et al.* (2019) — generate revealed-preference evidence against which transferred and stated-preference estimates can be validated. Fielding such studies for the Indian market, ideally in combination with a nationally representative incentive-compatible WTA survey, is accordingly identified as a priority alongside the data-collection agenda set out in Section 8.

The time-use valuation approach treats consumer time allocation as a revealed preference indicator, valuing time spent on free services at some fraction of its opportunity cost. Grounded in Becker's (1965) household production framework and applied by Goolsbee and Klenow (2006) and Brynjolfsson and Oh (2012), this approach faces the challenge that not all time spent on digital services generates positive surplus — habitual or addictive usage may be welfare-reducing (Allcott *et al.*, 2020).

A third approach uses advertising revenue as an observable calibration benchmark for advertising-supported digital services. Nakamura *et al.* (2016, 2017) use advertising-supported free content within an experimental production-account framework, while subsequent welfare studies compare platform revenues with consumer valuations. Advertising revenue should not, however, be interpreted as a theoretically established lower bound on consumer surplus. In a two-sided platform market, advertising revenue reflects advertisers' willingness to pay for access to user attention as well as platform market structure and monetisation strategies; its relationship to consumer welfare is therefore indirect. In this paper advertising revenue is consequently used only as a benchmark against which WTA- and time-use-based estimates can be compared.

The broader mismeasurement debate includes Feldstein (2017), who argued that real GDP growth systematically understates welfare improvement by failing to capture quality gains and new goods, and Syverson (2017), who concluded that while unmeasured digital output alone cannot explain the productivity slowdown, the mismeasurement is economically significant.

An important methodological issue concerns the aggregation of WTA when the available evidence provides service-specific median valuations rather than the full distribution of individual valuations. Multiplying a median WTA by the relevant user population does not, in general, recover the population sum of individual welfare. The empirical estimates in this paper are therefore interpreted as *median-equivalent welfare aggregates*, and the sensitivity of the results to alternative valuation assumptions is examined in Section 9. This distinction is particularly important given the absence of nationally representative Indian WTA microdata.

2.3. Evidence relevant to India

Direct empirical evidence on the willingness of Indian consumers to forgo free digital services remains limited, and we are not aware of a nationally representative Indian study using incentive-compatible choice experiments comparable to those employed in the international literature. Consequently, the present paper does not treat its WTA inputs as directly observed Indian valuations. Instead, it uses international experimental evidence, principally Brynjolfsson *et al.* (2019), as a starting point and transfers these values to the Indian context using purchasing-power and income adjustments. The limitations of such benefit transfer are important: preferences, income distributions, network effects, availability of offline substitutes and the composition of digital use can differ substantially between countries. The Indian estimates reported below should therefore be interpreted as experimental estimates designed to establish plausible orders of magnitude. A nationally representative Indian valuation experiment remains necessary before these estimates could be treated as statistical measures suitable for regular compilation.

3. India's free digital services ecosystem

3.1. Scale and scope

By March 2024, the Telecom Regulatory Authority of India (TRAI) reported 954.40 million internet subscriptions (TRAI, 2024). However, since many individuals hold multiple subscriptions, the number of unique individual internet users is substantially lower. The Internet and Mobile Association of India (IAMAI)-Kantar ICUBE 2024 report estimates approximately 820 million unique active internet users, a figure we adopt throughout this paper. The distinction among internet subscriptions, unique internet users, monthly active users (MAU), and daily active users (DAU) is important: these denominators cannot be interchanged, and service-specific MAU figures in Table 1 reflect the subset of users actively using each specific service. Usage is heavily skewed toward free, advertising-supported services: paid digital subscriptions (excluding telecommunications access) average only $3.2 per user per year, compared with $142 in the United States (Data.ai, 2024).

Table 1 presents the consolidated ecosystem of major free digital service categories.

Table 1: India's free digital services ecosystem, 2023–24

Service Category	MAU (million)	Daily Minutes per User	Annual Person-Hours (billion)	Indicative Ad-Revenue Intensity (₹per active user/year; non-additive)
(1)	(2)	(3)	(4)	(5)
Search engines	650	18	71.2	620
Video streaming (free tier)	520	48	152.0	540
Social media	530	42	135.5	480
Messaging	550	35	117.2	95
Email (free)	425	12	31.0	180
Maps/navigation	280	6	10.2	140
News/information	290	15	26.5	400
Educational platforms	180	8	8.8	220
Cloud storage (free tier)	310	4	7.5	65
Digital payments (Unified Payments Interface, UPI)	350	5	10.7	Nil
All Services/Unique User Base	**820**	**114** (MAU-weighted)	**570** (gross service-use hours)	**620** (= ₹50,800 cr ÷ 820M]

Notes: MAU = Monthly Active Users; UPI = Unified Payments Interface for each service category. Since individual users typically use multiple services, the service-specific MAU figures overlap substantially (the same user may appear in several rows). Column (3) reports average daily minutes for users of that specific service. Column (4) computes

service-specific annual person-hours as MAU $\times$ daily minutes $\times 365 \div 60$.

Service-specific MAU figures overlap because an individual may use several services. Such overlap does not by itself imply double counting of time: time spent on different services at different times remains distinct consumption time. Summing the service-specific person-hours produces approximately 570 billion gross service-use hours, equivalent to about 114 minutes per unique internet user per day when divided by the 820 million unique-user base. This figure should nevertheless be interpreted cautiously because some services may be used simultaneously—for example, navigation or messaging while consuming audio content—in which case gross service-use time exceeds elapsed clock time. The available data do not permit a systematic correction for such concurrent use.

Sources: IAMAI-Kantar ICUBE 2024; TRAI (2024); Data.ai 2024; Meta Q4 2023 Earnings; Alphabet FY2023 10-K; NPCI (2024); Comscore India 2024; Dentsu India, 2024.

Reconciliation note on Column (5). The service-specific figures in Column (5) are indicative advertising-intensity benchmarks compiled from heterogeneous platform disclosures and industry sources; they are not a mutually exclusive allocation of aggregate Indian digital advertising revenue across service categories. Differences in platform coverage, service definitions, geographic attribution and reporting basis mean that the row-level figures are not directly additive. Accordingly, multiplying each row by its service-specific MAU and summing across services produces a gross illustrative amount of approximately ₹1.28 lakh crore, which should not be interpreted as an estimate of total Indian digital advertising revenue. All calculations in this paper that require aggregate advertising revenue instead use the independently observed industry benchmark of ₹50,800 crore for free-to-consumer platforms (82 per cent of the ₹62,000 crore total reported by Dentsu India, 2024). The corresponding ₹620 figure in the final row is therefore calculated directly as ₹50,800 crore divided by 820 million unique active internet users and is not a weighted average of the service-specific entries in Column (5).

Several features merit emphasis. The weighted-average Indian internet user spends approximately 114 minutes daily on free digital services, computed as the MAU-weighted average across services. This represents a substantial reallocation of time, though less than the simple sum of service-specific minutes (193 minutes) because not all services are used by all users. The advertising revenue per user on free platforms (approximately ₹620/year, or roughly $7.50) is among the lowest globally for comparable engagement levels, suggesting that the gap between consumer valuation and platform revenue may be particularly large in India.

3.2. Usage heterogeneity

Aggregate usage averages conceal substantial differences in digital behaviour across population groups. Digital access, service composition and intensity of use vary by location, age, gender, employment status and household income. These differences matter for welfare estimation because both the opportunity cost of time and the value attached to individual digital services are likely to differ systematically across users. However, the available sources do not provide sufficiently harmonised demographic time-use estimates that are directly comparable with the service-specific MAU data in Table 1. We therefore do not attempt to infer demographic differences in welfare directly from aggregate usage statistics. Instead,

demographic variation enters the time-use sensitivity analysis through alternative wage and opportunity-cost assumptions. Developing a nationally representative digital time-use module that records service category, duration, concurrent use and socioeconomic characteristics would substantially improve future estimates.

4. Estimation framework and results

We employ three complementary estimation strategies: willingness-to-accept (WTA), time-use valuation, and an advertising-revenue calibration benchmark. While these approaches rest on different identifying assumptions, they share common user-base inputs, limiting their statistical independence (see Section 4.1). Convergence across methods provides useful corroboration but should not be interpreted as convergence of fully independent estimates.

4.1. Common framework and data quality

Notation. Subscript s indexes service categories ($s = 1, \ldots, 10$ per Table 1), subscript i indexes users, and subscript d indexes demographic cells. CS denotes consumer surplus with superscripts WTA, TIME, and AD distinguishing methods. All monetary values are in 2023–24 rupees. We use "lakh crore" ($= 10^{12}$ rupees) following Indian convention.

Shared data inputs. All methods draw on common user base inputs from TRAI (2024) and IAMAI-Kantar ICUBE 2024. These shared inputs are treated as uncertain in the Monte Carlo simulation, but errors in user base estimation affect all three methods in the same direction, limiting the degree of true statistical independence across methods.

Data quality. Our empirical foundation relies substantially on commercial industry bodies and platform self-disclosures, whose proprietary methodologies are not independently auditable. Where possible, we cross-validate against regulatory data (TRAI, the National Payments Corporation of India (NPCI), and the Reserve Bank of India (RBI)). Time-spent estimates, derived from panel-based measurement, may suffer from selection bias and measurement error. These uncertainties are partially but not fully captured in our Monte Carlo distributions; systematic bias across sources would shift results in ways not reflected in our 90 percent parameter-uncertainty intervals.

Monte Carlo convergence. Five independent replications with different random seeds confirmed that all reported percentiles were stable to within ± 0.1 lakh crore; increasing draws to 50,000 produced no material change.

4.2. Method 1: Willingness-to-accept valuation

For service s and user i, let W_{is} denote raw monthly WTA for continued access to the service. In the absence of Indian microdata on the distribution of W_{is}, the empirical analysis uses the transferred service-specific median WTA, denoted by $\widetilde{W}_s$. The corresponding annual median-equivalent welfare measure is:

$$CS^{WTA} = \sum_{s=1}^{S} N_s \times 12 \times \widetilde{W}_s$$

where $N_s$denotes the relevant number of users of service s. Because the underlying Indian distribution of individual WTA is not observed, multiplying the service-specific median by the number of users does not, in general, recover the population sum of individual valuations. We therefore describe this quantity as a *median-equivalent welfare aggregate*. It represents the aggregate that would result if each user were assigned the service-specific median valuation and should not be interpreted as a distribution-exact estimate of aggregate consumer surplus.

If the distribution of individual WTA were observed, aggregate welfare would instead be based on its population mean:

$CS^{WTA} = \sum_{s=1}^{S} N_s \times 12 \times E(W_s)$.

For example, if individual WTA is log-normally distributed with median $\widetilde{W}_s$ and log-standard deviation σ_s

$$E\left(W_s\right) = \widetilde{W}_s \exp\left(\frac{\sigma_S^2}{2}\right).$$

Reliable Indian estimates of σ_s are presently unavailable. We, therefore, retain the median-equivalent measure as the principal experimental specification and explicitly recognise this limitation. Future Indian WTA microdata would permit direct aggregation across respondents and eliminate this approximation.

In the absence of a directly estimated Indian σ_s, we can nonetheless bound the plausible understatement from median-based aggregation by borrowing dispersion estimates from the broader experimental-valuation literature, in the same spirit as the international parameter transfer used elsewhere in this paper. Stated WTA/WTP distributions surveyed in the hypothetical-bias meta-analyses underlying our β correction (List and Gallet, 2001; Murphy *et al.*, 2005) typically exhibit coefficients of variation (CV) in the region of 0.8 to 1.5. Under a log-normal assumption, CV relates to the dispersion parameter via $\sigma^2 = ln(1 + CV^2)$, and the mean-to-median ratio is $exp(\sigma^2/2)$; this yields ratios of approximately 1.28 at CV = 0.8 and 1.80 at CV = 1.5. Applying these ratios to the baseline median-equivalent estimate of ₹10.4 lakh crore implies a mean-based estimate of approximately ₹13.3–18.7 lakh crore — that is, correcting for right-skewness would, if anything, raise rather than lower the aggregate welfare estimate, since a mean necessarily exceeds a median under right skew. We emphasize that these CV values are illustrative bounds transferred from the general valuation literature rather than parameters estimated from Indian microdata, and we do not adopt the resulting mean-based range as an alternative headline estimate; the median-equivalent specification is retained as the principal measure precisely because no defensible India-specific dispersion parameter presently exists, and imposing an unvalidated distributional shape on Indian data we do not have would substitute one unverifiable assumption for another. The exercise is reported to characterize the direction and approximate scale of the aggregation bias rather than to correct for it, and it reinforces rather than weakens the case for prioritizing a nationally representative Indian WTA survey capable of directly recovering the shape of the underlying valuation distribution.

Raw WTA values cannot be used directly because they are subject to three well-documented biases, each of which inflates the estimate. We apply multiplicative correction factors to obtain adjusted surplus:

$CS^{WTA}_{adjusted} = \sum_{s=1}^{S} N_s \times 12 \times \frac{\widetilde{W}_s}{\beta} \times \delta \times \gamma_s.$

Here $\beta \sim U(2.0, 3.0)$ captures uncertainty concerning hypothetical-response inflation, while $\delta \sim U(0.60, 0.80)$ allows for substitution toward alternative services following the withdrawal of a particular service. The term γ_s is used only where an additional adjustment is required to reconcile the income level represented by the transferred valuation with that of the target Indian user population. It may be written as

$$\gamma_s = \left(\frac{Y_I}{Y_B}\right)^{\eta}$$

where Y_I is the relevant purchasing-power-parity (PPP)-adjusted income of the target Indian population, Y_B is the corresponding income level underlying the transferred valuation, and η is the assumed income elasticity of WTA. Importantly, γ_s is applied only when the transferred WTA has not already been adjusted to the target Indian income level. Where such adjustment is already embodied in the transferred value, $\gamma_s = 1$, thereby avoiding double adjustment for income differences.

4.3. Method 2: Time-use valuation

Following Becker (1965), when a consumer with a positive opportunity cost of time chooses to spend time on a free digital service rather than an alternative activity, their choice reveals that the value derived exceeds the opportunity cost of the time expended. Consumer surplus is the gap between value received and time cost incurred.

For a user in demographic cell d, let t_d denote daily hours devoted to free digital services and w_d the relevant market wage. We define the effective opportunity cost of time as:

$$w^*_d = \lambda \; w_d$$

where $\lambda \sim U(0.30, 0.50)$ recognises that most digital-service use displaces leisure or household time rather than paid working time.

For individuals without observed market earnings, the reference wage is first imputed as:

$$w_d = \xi \; w^{Cell}_d$$

where w^{Cell}_d is the median wage of the relevant demographic cell and $\xi \sim U(0.40, 0.80)$.

Consumer surplus associated with digital time is then approximated by:

$$CS^{TIME} = 365 \; \phi\mu \sum_{d}^{D} N_d w^*_d t_d$$

where D denotes the number of demographic cells, N_d is the number of users in cell d and t_d is mean daily time devoted to the relevant free digital services. The parameter $\phi \sim U(0.25, 0.75)$ represents net surplus relative to the effective opportunity cost of time, while $\mu \sim Triangular(0.40, 0.60, 0.80)$ discounts digital time that substitutes for comparably valued offline activity. Given the limited direct Indian evidence for these parameters, the time-use calculation is treated as an order-of-magnitude robustness benchmark rather than as an independently identified estimate of consumer welfare.

4.4. Advertising-revenue calibration benchmark

Platforms provide free services funded by advertising. The empirical literature has explored the relationship between per-user advertising revenue and consumer valuations. Varian (2018) estimates that consumer surplus from Google Search is approximately 8–12 times the advertising revenue Google earns per user. Brynjolfsson *et al.* (2019) imply a ratio of approximately 16 for Facebook based on their WTA estimates relative to Meta's per-user advertising revenue.

We use this relationship as a calibration benchmark:

$$CS^{AD} = \theta \times R_{total}$$

where $R_{total} =$ ₹50,800 crore is total digital advertising revenue from free-to-consumer platforms in 2023–24 (82 percent of ₹62,000 crore total; Dentsu India, 2024), and $\theta \sim U(8, 20)$.

We emphasize that this is a calibration exercise rather than an independently identified consumer-surplus estimator. In a two-sided platform market, advertising revenue is determined by advertisers' willingness to pay for user attention and by platform market structure. There is no general theoretical result establishing that consumer surplus must exceed advertising revenue; the empirical regularity documented by Varian and Brynjolfsson *et al.* is contingent on observed market structures and user valuations. For India, where per-user advertising revenue is among the lowest globally (approximately ₹620/year), the multiplier may be particularly high because platforms provide similar-quality services despite earning far less per user, but this remains an empirical conjecture. It should be noted that advertising revenue is already part of measured GDP: as discussed in Section 2.1, it is recorded either as intermediate consumption by advertisers or as the platform's own output of advertising services, and is therefore fully captured within the core production boundary. The point at issue here is different — namely, whether observed advertising revenue provides a reliable empirical multiplier for inferring the consumer surplus that households derive from free access, a welfare quantity that lies outside GDP by construction.

Nakamura *et al.* (2017), often cited in this context, are primarily concerned with production-account treatment of advertising-supported free content rather than establishing a general advertising-revenue multiplier for consumer surplus.

This method also cannot capture surplus from non-advertising-funded services such as UPI, which we estimate contributes approximately 9 percent of total consumer surplus through the WTA and time-use methods.

4.5. Monte-Carlo combination

We treat the bias-adjusted WTA specification as the principal experimental welfare measure because it is most directly related to consumers' valuations of access. The time-use and advertising-revenue calculations are reported as calibration and robustness exercises. For completeness, we also report an equal-weight average of the three measures, but this combined statistic has no optimal-weighting interpretation and is not used as the paper's headline estimate. In particular, convergence among the three approaches should not be interpreted as statistical validation because the methods share several data inputs and are

affected by common sources of model uncertainty

Each of the three methods depends on parameters that are uncertain — in some cases substantially so. Rather than presenting a single point estimate with an ad hoc sensitivity range, we propagate uncertainty from all input parameters simultaneously using Monte Carlo simulation. This produces a full probability distribution over consumer surplus estimates, from which we can extract 90 percent parameter-uncertainty intervals that honestly characterise what we know and do not know.

Procedure. We draw 10,000 independent realisations of the full parameter vector. For each draw ($k = 1, .., 10,000$):

(a) Sample all uncertain parameters from their specified distributions (see Table 2 below).

(b) Compute the resulting surplus estimate under each of the three methods independently: CS_k^{WTA}, CS_k^{Time}, and CS_k^{AD}.

(c) Record all three estimates.

For each of 10,000 independent draws of the full parameter vector (Table 2), we compute all three method-specific estimates. We designate the bias-corrected WTA estimate as our principal welfare estimator, given its direct theoretical grounding in Hicksian equivalent variation, and present the time-use and advertising-revenue estimates as robustness benchmarks.

For comparison purposes, we also report a combined estimate using equal-weighted averaging:

$$CS_k^{COMBINED} = \tfrac{1}{3}\left(CS_k^{WTA} + CS_k^{TIME} + CS_k^{AD}\right).$$

Equal weighting is a convenience rather than a statistically justified aggregation: the three methods are not repeated measurements of exactly the same quantity with known sampling properties, and their identifying assumptions differ substantially. The combined estimate should be interpreted as a summary indicator rather than a precision-weighted optimal estimate. As robustness checks, alternative weights of (0.50, 0.30, 0.20) — privileging the WTA method — and (0.25, 0.50, 0.25) yield median combined estimates of ₹10.0 and ₹10.5 lakh crore respectively.

We refer to the 5^{th}–95^{th} percentile range of the Monte Carlo output distribution as a "90 percent simulation-based uncertainty interval." This terminology reflects the fact that the interval propagates *assumed* parametric uncertainty through the model; it is not a Bayesian credible interval derived from a formally estimated posterior distribution. The interval honestly characterizes Phe sensitivity of results to parameter assumptions but does not capture systematic biases (such as demand-curve shape assumptions) that affect all draws equally.

Table 2 documents the distributional assumptions for each stochastic parameter, the empirical basis for each distribution choice, and the method(s) in which each parameter enters.

Table 2: Monte Carlo parameter distributions

Parameter (1)	Distribution (2)	Basis (3)	Method(s) (4)
WTA per user, by service	Transferred service-specific median; no individual-level parametric distribution imposed.	International experimental evidence transferred and adjusted to the Indian context; alternative valuation levels examined in sensitivity analysis	WTA
Hypothetical bias correction (β)	Uniform(2.0, 3.0)	List and Gallet (2001); Murphy *et al.* (2005)	WTA
Substitution discount (δ)	Uniform(0.60, 0.80)	Brynjolfsson *et al.* (2019) cross-service experiments	WTA
Active user count	Triangular(DAU at 67% of MAU; MAU; MAU + 5%)	Platform disclosures; TRAI	WTA; Time-use
Income elasticity of WTA (η)	Triangular(0.3, 0.5, 0.7)	Cross-country income variation implicit in Brynjolfsson *et al.* (2019); consistent with digital-goods income-elasticity estimates in the literature (Coyle and Nguyen, 2020)	WTA
Surplus ratio (ϕ)	Uniform(0.25, 0.75)	Allcott *et al.* (2020) deactivation experiment bounds	Time-use
Wage rate	Periodic Labour Force Survey (PLFS) 2023–24 distribution by demographic cell	National Statistical Office (NSO) earnings data;	Time-use
Digital attribution factor (μ)	Triangular(0.40, 0.60, 0.80)	Substitution rates for offline leisure displacement	Time-use
Ad revenue multiplier (θ)	Uniform(8, 20)	Brynjolfsson *et al.* (2019); Varian (2018)	Advertising
Non-worker discount (ξ)	Uniform(0.40, 0.80)	Abraham and Mackie (2005); Bridgman (2016)	Time-use
Leisure-time discount (λ)	Uniform(0.30, 0.50)	Opportunity cost of non-work time; sensitivity assumption	Time-use

Note:

The three methods share common user base inputs (TRAI, IAMAI), limiting true independence. Cross-method convergence provides less corroboration than fully disjoint methods would.

Table 3: Experimental welfare estimates and calibration benchmarks, 2023–24 (₹ lakh crore)

Method	5th %tile	25th %tile	Median	75th %tile	95th %tile	IQR
(1)	(2)	(3)	(4)	(5)	(6)	(7)
WTA-based median-equivalent welfare measure (principal estimate)	5.8	8.2	10.4	13.8	19.6	5.6
Time-use valuation	7.2	10.8	14.1	18.3	24.8	7.5
Advertising-revenue calibration benchmark	4.4	5.6	7.1	8.6	9.9	3.80
Equal-weight combined diagnostic	**5.8**	**7.8**	**9.6**	**12.2**	**16.4**	**4.4**
Combined as % of GDP (₹302 lakh crore)	1.9%	2.6%	3.2%	4.0%	5.4%	1.4 pp

Notes: All monetary values in 2023–24 current prices. The 90 percent simulation-based uncertainty interval for the combined estimate is ₹5.8–16.4 lakh crore. The WTA-based estimate is our preferred principal estimate; the combined figure is reported for comparison.

4.6. Results

Table 3 presents the Monte Carlo results for each method individually and for the combined estimate.

As an internal consistency check, the advertising-revenue calibration has a particularly transparent analytical range. With aggregate advertising revenue fixed at ₹50,800 crore and $\theta \sim U(8,20)$, the implied welfare benchmark is bounded between ₹4.064 lakh crore and ₹10.16 lakh crore. The corresponding 5th, 25th, 50th, 75th and 95th percentiles are approximately ₹4.4, ₹5.6, ₹7.1, ₹8.6 and ₹9.9 lakh crore, respectively. These analytically implied values provide a direct check on the Monte Carlo implementation.

The WTA-based measure is treated as the principal experimental welfare estimate because it most directly addresses the value consumers attach to continued access to the services considered. The time-use estimate and advertising-revenue calculation serve different analytical purposes and are reported primarily as robustness and calibration benchmarks. Their numerical proximity to, or divergence from, the WTA estimate should not be interpreted as statistical validation because the approaches rely on different identifying assumptions and share some common inputs. The width of the WTA simulation interval reflects uncertainty in the assumed parameters and transferred valuations; it should therefore be interpreted as a model- and parameter-uncertainty range rather than as a conventional sampling-based confidence interval.

Before presenting the results, a conceptual caveat is warranted. The WTA method as applied here measures total access value — the compensation required for complete loss of access — which includes quasi-rents from network effects rather than conventional Marshallian surplus measured along a demand curve. For network-dependent services such as WhatsApp, losing access would impose coordination costs far exceeding the hedonic value

of messaging functionality alone. The hypothetical bias correction partially compensates for this but is calibrated for hypothetical inflation, not for the conceptual gap between access value and Marshallian surplus.

4.7. Service composition of the WTA-based welfare estimate

To maintain consistency with the designation of WTA as the principal welfare estimator, the service-level decomposition is based on the median-equivalent WTA specification. For service s, its share of the aggregate is proportional to $N_s W_s \gamma_s$, since the common hypothetical-bias correction (β), substitution discount (δ), and annualization factor cancel when contributions are expressed as shares. Table 4 makes the arithmetic underlying the reported decomposition explicit and, in addition to the service shares, reports the implied per-user valuations required to reproduce the aggregate estimate..

For transparency, the service shares are obtained by normalizing the service-specific contributions entering the WTA calculation. Specifically, for service s, the relevant contribution is proportional to $N_s W_s \gamma_s$, because the common hypothetical-bias correction β, substitution factor δ, and annualisation factor cancel when service contributions are expressed as shares of the total. Thus,

$$Share_s = \frac{N_s \widetilde{W}_s \gamma_s}{\sum_{j=1}^{S} N_j \widetilde{W}_j \gamma_j} \ .$$

Replication of the decomposition therefore requires disclosure not only of the service-specific user counts but also of the transferred WTA values, their original sources and the income/PPP adjustments applied to them. Table 4 reports these inputs explicitly so that the service composition can be independently reproduced.

The implied adjusted annual welfare per user is the service-specific welfare contribution divided by the corresponding MAU, and the monthly figure is its annual value divided by 12. The final column reports the combined transferred-valuation and income-adjustment term $\widetilde{W}_s \gamma_s$ implied at the baseline common correction factors $\beta = 2.5$ and $\delta = 0.70$. Because the underlying foreign valuation and the India adjustment cannot be separately identified from the information presently available, the table reports their product rather than assigning an arbitrary decomposition between W_s and γ_s. Figures may not sum exactly because of rounding.

Table 4 shows that the reported decomposition implies substantial variation in adjusted welfare per user across service categories. The implied annual values range from approximately ₹335 per cloud-storage user to around ₹3,600 per video-streaming user, with search, messaging and video generating the highest per-user values. These figures should be interpreted as values implied by the transferred-WTA specification rather than as directly observed Indian WTA estimates. Their principal purpose is to make the service decomposition transparent and reproducible and to indicate the service-level valuations that future Indian incentive-compatible experiments would need to test.

Under the transferred WTA specification, search engines account for the largest estimated share (approximately 22 per cent), followed by messaging (19 per cent), video streaming (18 per cent), social media (16 per cent) and UPI digital payments (9 per cent). Maps account for approximately 6 per cent, email 4 per cent, education 3 per cent, news 2 per cent

Table 4: Service-level replication of the WTA-based welfare decomposition, 2023–24

Service	MAU (million)	Reported welfare share (%)	Welfare contribution (₹lakh crore)	Implied adjusted welfare per user/year (₹)	Implied adjusted WTA per user/month (₹)	Implied $W_s\gamma_s$ before common corrections (₹/month)
Search engines	650	22	2.288	3,520	293	1,048
Messaging	550	19	1.976	3,593	299	1,069
Video streaming	520	18	1.872	3,600	300	1,071
Social media	530	16	1.664	3,140	262	934
UPI/digital payments	350	9	0.936	2,674	223	796
Maps/ navigation	280	6	0.624	2,229	186	663
Email	425	4	0.416	979	82	291
Educational platforms	180	3	0.312	1,733	144	516
News/ information	290	2	0.208	717	60	213
Cloud storage	310	1	0.104	335	28	100
Total	—	**100**	**10.400**	—	—	—

Notes: MAU figures are from Table 1. Service-specific welfare contributions are obtained by applying the reported WTA shares to the principal median-equivalent welfare estimate of ₹10.4 lakh crore.

and cloud storage 1 per cent. These shares reflect the combined effect of service-specific user populations and adjusted transferred WTA values and should not be interpreted as precisely estimated structural welfare weights. Particularly noteworthy is UPI: despite generating no advertising revenue in the calibration used here, it receives a positive WTA-based welfare contribution. This illustrates why advertising revenue cannot provide a comprehensive allocation of consumer welfare across zero-price digital services.

5. Situating estimates within the 2025 SNA framework

The 2025 SNA represents the most significant revision to international statistical standards since the 2008 SNA. Its treatment of digitalization, set out principally in Chapter 22, is largely enabling rather than prescriptive — establishing frameworks within which national statistical offices may develop approaches rather than mandating specific imputations. Three features are directly relevant to this paper, and it is essential to distinguish clearly among the concepts they address.

5.1. Three distinct concepts

Throughout this paper we maintain a strict distinction among three concepts that are easily conflated:

(a) ***Core SNA GDP*** — the production-boundary aggregate, in which free platform services do not generate recorded household consumption expenditure. Consumer surplus is excluded by design, not by oversight.

(b) ***2025 SNA extended account for free platform services*** — Chapter 22 of the 2025 SNA encourages the construction of extended or thematic accounts in which household consumption of free platform services is made visible. Crucially, for this extended account, the SNA regards the *platform's cost of producing the free services* as the most suitable valuation method, not consumer willingness-to-pay or advertising revenue per user. Furthermore, the SNA explicitly notes that recording these services as household consumption while simultaneously retaining their existing use as inputs to advertising and data production would create double counting within GDP. Extended accounts therefore, require careful supply-use rebalancing.

(c) ***Consumer surplus / welfare measures outside GDP*** — the 2025 SNA endorses a "Beyond GDP" measurement agenda as a legitimate complement to production-based accounting. Consumer surplus from free digital services is a welfare indicator that belongs in this category. Our estimate of ₹10.4 lakh crore is not "missing GDP" but a welfare measure that GDP is not designed to capture.

5.2. Data as a produced asset

The 2025 SNA extends the asset boundary to recognize data as a produced asset (addressed in the asset classification chapters of the 2025 SNA). However, the SNA draws an important distinction between the platform's production of a data asset through collection, curation, and processing, and the household's passive permission for its characteristics or behaviour to be observed. Passive permission to collect household data is not treated as household production under the SNA framework. The SNA allows that the free platform service may be regarded, in an extended account, as payment in kind for a licence to collect data, with the household-side receipt classified as rent rather than as compensation for production. Only where households actively perform actions specifically assisting data collection — beyond normal service usage — might household production arise. Chapter 22 of the 2025 SNA provides this distinction explicitly.

The platform produces the data asset; the household provides access to observable behaviour, for which the free service may constitute implicit compensation (a licence fee or rental payment in the extended account).

5.3. Thematic or extended accounts

The 2025 SNA's framework for satellite and thematic accounts is set out in Chapter 38. Digital-economy thematic accounts may be constructed at multiple levels of extension. The least expansive level captures ICT-sector output within the existing production boundary; a more extended level may include an imputation for household consumption of free

Table 5: Treatment of free digital services under the 2025 SNA

Component	Correct 2025 SNA treatment	Current Indian recording	Measurement gap
(1)	(2)	(3)	(4)
Platform output (advertising services)	Core production boundary: market output (Chapters 6–7)	Recorded as IT/advertising output	Minor (cross-border attribution)
Data produced by platforms from user observation	Asset boundary: data as produced asset (asset classification chapters)	Not separately identified	Platform-side data asset not capitalised
Household permission for data collection	Extended account: implicit rental income to household (Ch. 22)	Not recorded	Licence-fee / rental flow unmeasured
Free services consumed by households	Extended account (Ch. 22, Ch. 38): valued at producer cost	Not recorded as household consumption	Full consumption value unmeasured; cost-based valuation preferred by SNA
Consumer surplus from free services	Beyond GDP welfare indicator	Not measured; excluded by design	₹10.4 lakh crore (this paper)

Sources: 2025 SNA, Chapters 6, 7, 22, and 38; BPM7 (IMF, 2025); OECD (2020); authors' assessment of Indian national accounts recording practices based on NSO (2024) methodology documentation.

services, valued at producer cost. Chapter 22 provides guidance on these extensions.

5.4. Mapping our estimates to the 2025 SNA

Table 5 maps the components of our analysis onto these three concepts.

The most immediately actionable gap for India's NSO is the construction of a digital-economy satellite account along the lines encouraged by Chapter 22. Within this account, household consumption of free platform services would be made visible, valued at the platform's cost of producing those services. A rough order of magnitude can be obtained from platform cost data, though obtaining service-level cost allocations from multinational platforms operating in India will present practical challenges.

It is important to note that valuing free services at advertising revenue per user (as sometimes proposed in the literature) is not the preferred 2025 SNA valuation for the extended account. Furthermore, simply adding imputed free-service consumption to GDP without rebalancing the supply-use framework would create double counting, since the resources used to produce free services are already recorded as costs of producing advertising output.

Our consumer-surplus estimates (Section 4) are explicitly welfare measures belonging

Table 6: Alternative measurement perspectives on free digital services

Measurement perspective	Valuation	Status under 2025 SNA	Interpretation
(1)	(2)	(3)	(4)
Core SNA accounts	Market / accounting prices	Core accounts	Records platform market output but excludes consumer surplus
Extended account for free platform services	Producer cost	Encouraged supplementary/extended treatment	Makes household use of free services visible without treating WTA as GDP
Consumer-welfare measure	WTA/consumer valuation	Outside core GDP	Measures welfare associated with access to free services
Advertising-revenue calibration	Advertising revenue/multiplier	Analytical benchmark only	Robustness comparison; not SNA-preferred valuation

to the "Beyond GDP" framework. They complement, but do not substitute for, the cost-valued extended account.

The preceding discussion suggests that free digital services can be viewed from several distinct measurement perspectives, each addressing a different analytical question and requiring a different valuation basis. To make these distinctions explicit, Table 5 summarises the relationship among the treatment of free digital services in the core SNA accounts, their representation in an extended account, and their valuation for consumer-welfare analysis. The table also identifies the more limited role of advertising revenue as an analytical calibration benchmark. These approaches should not be interpreted as alternative estimates of the same aggregate; rather, they correspond to different measurement objectives and accounting boundaries.

The central implication of Table 5 is that these measures answer different questions and should not be added mechanically to one another. The cost-valued extended account concerns the organization and visibility of production and consumption within an expanded accounting framework, whereas WTA-based consumer surplus concerns welfare. The latter is therefore complementary to, rather than a correction of, GDP.

As Table 6 makes clear, the distinction between production measurement and welfare measurement is central to the interpretation of the estimates presented in this paper. Core GDP records economic production according to SNA accounting principles and excludes consumer surplus by design. The extended-account approach recommended in the 2025 SNA can make household consumption of free platform services more visible, but values such services on a producer-cost basis rather than by consumers' willingness to pay or accept. The WTA-based estimates developed in this paper address a different question: the welfare consumers associate with continued access to these services. Advertising revenue provides a useful observable benchmark but is neither the SNA-preferred valuation for the extended account nor a direct measure of consumer surplus. Accordingly, the four perspectives in Table 5 should be reported alongside one another rather than combined into a single augmented

GDP measure.

6. Implications for welfare-augmented growth measurement

The final distinction in Table 5—between production-based accounting and consumer-welfare measurement—also provides the appropriate starting point for considering welfare-augmented growth measures such as GDP-B. Whereas the WTA estimates reported in Section 4 describe the level of consumer valuation in a particular period, a welfare-growth measure requires information on how that valuation changes over time.

Brynjolfsson *et al.* (2025) develop GDP-B as a framework for incorporating the welfare contributions of new and free goods into measures that complement conventional production-based statistics. The framework is particularly relevant to rapidly expanding digital economies because changes in the consumer valuation of zero-price services need not be reflected proportionately in market expenditure.

Applying GDP-B rigorously to India requires repeated observations on consumer valuations over time. In particular, the welfare contribution of free digital services to growth depends on changes in consumer surplus rather than its level:

$$\Delta CS_t = CS_t - CS_{t-1} \ .$$

A valid implementation would therefore require comparable annual or periodic estimates of WTA, appropriately adjusted for changes in the number and composition of users, the introduction and disappearance of services, quality change and inflation. The cross-sectional 2023–24 estimates developed in this paper do not provide this longitudinal information.

We, therefore, refrain from constructing a numerical GDP-B growth series from interpolated consumer-surplus levels. Instead, our results establish a benchmark welfare level for 2023–24 that could serve as a starting point for future longitudinal measurement. Repeated incentive-compatible WTA experiments—preferably conducted at regular intervals using a consistent survey design—would permit India to construct a genuine welfare-growth series and assess the contribution of free digital innovation using the GDP-B methodology.

This distinction is important. A large level of consumer surplus does not necessarily imply a correspondingly large annual welfare-growth contribution: GDP-B growth depends on the change in welfare between periods. As internet penetration approaches saturation, extensive-margin gains may diminish even while quality improvements, new services and artificial intelligence generate continuing intensive-margin welfare gains.

7. International context

Evidence from other economies consistently indicates that zero-price digital services can generate consumer valuations substantially exceeding the market expenditures directly associated with their use. Experimental studies for the United States and other advanced economies have demonstrated large valuations for services including search, social media, digital maps and online communication. However, the magnitudes reported across studies are not directly comparable because survey designs, service definitions, elicitation methods,

user populations and price concepts differ.

These comparability problems are particularly important when transferring results to India. Lower per-capita income might reduce willingness to pay or accept in monetary terms, while extensive reliance on free rather than subscription-based services, strong network effects and limited availability of some offline substitutes may work in the opposite direction. Consequently, cross-country ratios of estimated digital consumer surplus to GDP should not be interpreted as comparable national-accounting statistics.

The international literature therefore supports a qualitative rather than a precise quantitative conclusion: substantial consumer welfare from zero-price digital services is a general feature of digital economies and not unique to India, but its magnitude cannot be read off from other countries' experience and must ultimately be estimated using country-specific evidence. This observation motivated the measurement framework developed in Section 4 of this paper, and it is revisited in the Conclusion when we set out priorities for building India's own valuation evidence base.

8. Policy implications and implementation road map

Our results suggest a phased measurement agenda for India's National Statistical Office. In the near term, priority should be given to improving the visibility of digital production and transactions within supply-use and thematic-account frameworks and to developing information on the costs incurred by platforms in supplying zero-price services to Indian users. Such information would support the cost-of-production valuation preferred by the 2025 SNA for an extended account of free platform services. In parallel, experimental welfare measurement should be developed separately, beginning with nationally representative, incentive-compatible WTA surveys and improved digital time-use information. Over the longer term, repeated valuation exercises could support a broader welfare dashboard and permit the construction of longitudinal measures such as GDP-B. Maintaining the distinction between cost-valued extended accounts and consumer-valued welfare indicators is essential to preserving the conceptual integrity of both exercises.

Three broader implications merit mention. First, the potentially substantial consumer welfare indicated by our experimental estimates should be considered in regulatory cost-benefit analysis for competition policy and digital regulation affecting free service access. Second, the substantial consumer welfare generated by free digital services — much of it provided by platforms headquartered outside India — is relevant context for understanding the digital economy's contribution to Indian wellbeing. However, consumer surplus generated by foreign platforms is not directly evidence regarding the appropriate allocation of taxable profits; the taxation question requires separate analysis of value creation, profit allocation, and international tax frameworks. Third, the high surplus-to-cost ratio for publicly funded digital infrastructure — particularly UPI — provides a welfare-based framework for evaluating digital public goods investment.

9. Sensitivity analysis

The Monte Carlo analysis in Section 4 propagates uncertainty across the parameters entering the experimental estimates. Because the WTA-based median-equivalent measure is

Table 7: Sensitivity of experimental welfare estimates to key assumptions across valuation methods

Parameter	Baseline	Low (10th %tile)	CS at Low	High (90th %tile)	CS at High	Range
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Wage rate	PLFS median	PLFS 25th %ile	7.4	PLFS 75th %ile	12.1	4.7
Surplus ratio (ϕ)	0.50	0.30	7.8	0.70	11.4	3.6
Ad revenue multiplier (θ)	14	9.2	7.9	18.8	11.2	3.3
Hypothetical bias (β)	2.5×	2.1×	11.4	2.9×	8.2	3.2
Digital attribution (μ)	0.60	0.44	8.5	0.76	10.8	2.3
Active user count	820 mn	680 mn	8.0	860 mn	10.1	2.1
Substitution discount (δ)	0.70	0.62	8.7	0.78	10.6	1.9
Income elasticity (η)	0.50	0.35	8.9	0.65	10.4	1.5

Note: Each row varies one parameter while holding all other parameters at their baseline values. The parameters enter different component methods: β, δ, and η primarily affect the WTA specification; wage rate, ϕ, and μ affect the time-use specification; and θ affects the advertising-revenue calibration. Active-user assumptions affect the relevant user-based calculations. Accordingly, the table reports sensitivity of the combined diagnostic rather than of the WTA-based principal estimate.

designated as the paper's principal welfare estimator, the present section focuses primarily on the sensitivity of that measure to its key assumptions. Sensitivity of the time-use and advertising-revenue calculations is secondary, since these measures are used as robustness or calibration benchmarks rather than as alternative headline estimates.

Table 7 presents a structured one-at-a-time sensitivity analysis of the equal-weight combined diagnostic, varying selected parameters from the WTA, time-use and advertising-revenue approaches while holding all other parameters at their baseline values. For each parameter, Columns (3) and (5) report the low and high values examined, and Columns (4) and (6) report the resulting combined welfare estimate (₹ lakh crore) when that parameter alone is moved to the low or high value while every other parameter is held at its Table 2 baseline; Column (7) reports the resulting range (High minus Low). Reading down Column (7) therefore, ranks parameters by how much the combined estimate moves in response to each one, from the wage rate at the top (the most influential assumption) to income elasticity at the bottom (the least influential). The exercise is intended to identify the assumptions to which the cross-method summary measure is most sensitive and should not be interpreted as a sensitivity analysis of the principal WTA estimate alone.

Extended sensitivity to the substitution discount reveals that ($\delta = 0.40$) (high substitutability) and ($\delta = 0.95$) (near-zero substitutability) yield combined medians of ₹7.2 and ₹11.8 lakh crore respectively — within the overall 90 percent parameter-uncertainty interval and not altering the qualitative finding.

9.1. Sensitivity to the aggregation of WTA

A further source of uncertainty concerns the aggregation of service-specific WTA values. Our principal calculation uses transferred median WTA values and therefore represents a median-equivalent welfare aggregate rather than the population sum of individual valuations. Without Indian microdata, the shape and dispersion of the WTA distribution cannot be identified reliably. Consequently, we do not attempt to impose a specific linear or other functional form on the underlying demand curve.

As an illustrative sensitivity exercise, we uniformly vary the transferred service-specific median WTA vector by ±30 per cent around the baseline specification. Holding the remaining WTA-model assumptions unchanged, a 30 per cent downward adjustment would reduce the baseline median-equivalent estimate of ₹10.4 lakh crore to approximately ₹7.3 lakh crore, while a 30 per cent upward adjustment would increase it to approximately ₹13.5 lakh crore. These values are mechanical sensitivity scenarios rather than statistical confidence limits or formal welfare bounds.

The exercise demonstrates that the magnitude of the estimated welfare value remains materially dependent on the transferred WTA inputs. It therefore reinforces the paper's central statistical recommendation: nationally representative, incentive-compatible Indian WTA experiments are the highest-priority data investment for producing more defensible consumer-welfare estimates of free digital services.

10. Limitations

Several limitations warrant acknowledgement, ordered by consequence for our central estimates.

First, and most importantly, the WTA inputs are not directly observed for India. They are transferred from international experimental evidence and adjusted to the Indian context. Neither PPP nor income adjustment can fully capture cross-country differences in preferences, network effects, available substitutes, digital literacy or service quality. Moreover, because Indian WTA microdata are unavailable, the paper uses median-equivalent rather than distribution-exact aggregation. These limitations imply that the reported monetary estimates should be interpreted as experimental orders of magnitude. A nationally representative Indian survey using incentive-compatible elicitation is therefore, the single most important requirement for improving the estimates.

Second, the time-use benchmark relies on strong assumptions concerning the opportunity cost and marginal welfare value of digital time. Time spent on digital services cannot automatically be equated with consumer surplus: some use may substitute for similarly valued offline leisure,

Third, substitution patterns are simplified. Our synergy discount (= 0.60 – 0.80) captures cross-service substitutability only approximately; a structural demand model with estimated cross-elasticities would improve precision but requires unavailable micro-data.

Fourth, the three methods are not fully independent: they share user base data and certain calibration parameters draw on overlapping empirical literature. Cross-method convergence is therefore, partially mechanical.

Fifth, and importantly, this paper's framework measures only the gross positive value that consumers derive from access to free digital services; it does not net out the welfare costs that accompany that access, nor does it address producer surplus or the distribution of net benefits across the population. Three distinct extensions would be required for a genuinely comprehensive welfare accounting:

- The first concerns digital "bads". Free digital services also generate documented negative externalities, including financial losses from cyber fraud and phishing, the erosion of privacy through data collection and secondary use, the mental health costs associated with problematic or compulsive use (particularly among adolescents), and the social costs of misinformation and manipulated content spread through social media and messaging platforms. Each of these has a growing dedicated empirical literature — for example, on the welfare costs of social media use (Allcott, Braghieri, Eichmeyer, and Gentzkow, 2020, cited above as Allcott *et al.*) and on privacy valuation more generally — but the magnitudes are not currently estimated on a basis comparable to the WTA framework used here, and Indian evidence is particularly sparse. Our ₹10.4 lakh crore estimate should therefore, be understood as a measure of gross access value, not of net welfare; the true net welfare contribution of free digital services is smaller, potentially substantially so, once these costs are netted out. Constructing a comparable "digital bads" valuation exercise — ideally using elicitation methods consistent with the WTA approach used here, so that gross and negative components can be legitimately netted against one another — is accordingly among the priorities we identify for future work in Section 8.
- The second concerns producer surplus. Platforms capture value through advertising revenue, data monetization and market power; a complete welfare accounting would set consumer surplus alongside producer surplus (and, where relevant, government revenue from digital taxation) to assess how the total surplus generated by the digital economy is divided among households, firms and the state. Our framework, being calibrated from the household side, is silent on this division.
- The third concerns distributional effects. The MAU-weighted averages underlying Table 1 and the aggregate estimates in Table 3 conceal potentially large differences in consumer surplus across income groups, urban and rural populations, age cohorts and gender, both because usage intensity varies systematically across these groups (Section 3.2) and because the opportunity cost of time and marginal utility of money used to value that usage are also likely to differ across them. A population-average welfare estimate of the kind reported here is therefore compatible with very different distributional profiles — for example, a pattern in which affluent, urban, digitally literate users capture a disproportionate share of estimated welfare would have materially different policy implications from a more evenly distributed pattern, even at an identical aggregate. Disaggregating the valuation exercise by demographic and socioeconomic strata, which would require the nationally representative time-use and WTA microdata called for throughout this paper, is necessary before firm distributional conclusions can be drawn.

Finally, the rapid emergence of generative AI services at zero price means these estimates reflect the 2023–24 landscape and may require significant revision as AI reshapes the free digital ecosystem.

11. Conclusion

Free digital services have become an important component of household consumption and everyday economic life in India, yet the welfare they generate is only imperfectly reflected in market-based economic statistics. As the international evidence reviewed in Section 7 makes clear, this is not a uniquely Indian phenomenon, but its magnitude for India could not be inferred from other countries' experience alone. This paper develops an experimental framework for assessing the order of magnitude of that welfare and, equally importantly, clarifies how such estimates should be related to the 2025 SNA.

Using transferred WTA evidence adjusted to the Indian context, the principal specification produces a median-equivalent welfare estimate of approximately ₹10.4 lakh crore in 2023–24, or about 3.4 per cent of nominal GDP. This figure should not be interpreted as missing GDP or as a statistical estimate of output omitted from India's national accounts. It is an experimental welfare measure whose precise magnitude remains sensitive to the transferred WTA values, assumptions concerning substitution and hypothetical response bias, and the absence of Indian micro-level valuation data. Time-use and advertising-revenue calculations provide useful benchmarks but do not eliminate these sources of uncertainty.

The 2025 SNA provides an important conceptual framework for separating these issues. As summarised in Table 5, core GDP remains a measure of production and deliberately excludes consumer surplus, while an extended account can make household consumption of free platform services more visible using producer cost as the preferred valuation basis. Consumer-surplus estimates answer a different question and are therefore appropriately treated as complementary welfare indicators rather than additions to GDP.

The most important next step is empirical rather than conceptual. India would benefit from nationally representative incentive-compatible WTA experiments, improved service-specific digital time-use statistics and information on the costs incurred by platforms in supplying free services to Indian users. Repeating such exercises over time would also permit application of welfare-growth frameworks such as GDP-B without relying on interpolated consumer-surplus series.

The central contribution of the paper is therefore not to propose a revised measure of Indian GDP, but to demonstrate how the economically important welfare generated by zero-price digital services can be measured experimentally while maintaining a clear boundary between production accounting and welfare analysis. That distinction is essential for extending economic measurement in the digital era without compromising the conceptual coherence of the national accounts. As generative AI services increasingly enter the zero-price digital ecosystem, the measurement challenges documented in this paper are likely to intensify, reinforcing the urgency of building the statistical infrastructure — nationally representative valuation experiments, platform cost data, and digital time-use surveys — that would allow India to track welfare changes in real time rather than relying on retrospective, transferred estimates.

Acknowledgements

Authors are grateful to the anonymous reviewer for a detailed and constructive comments, which has strengthened both the statistical transparency and the presentational clarity of the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abraham, K. G. and Mackie, C. (eds.) (2005). *Beyond the Market: Designing Nonmarket Accounts for the United States*. National Academies Press, Washington, D.C.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, **110**, 629–676.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, **75**, 493–517.

Bridgman, B. (2016). Home productivity. *Journal of Economic Dynamics and Control*, **71**, 60–76.

Brynjolfsson, E., Collis, A., and Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences*, **116**, 7250–7255.

Brynjolfsson, E. and Oh, J. (2012). The attention economy: Measuring the value of free digital services on the internet. *Proceedings of the International Conference on Information Systems*.

Brynjolfsson, E., Collis, A., Diewert, W. E., Eggers, F., and Fox, K. J. (2025). GDP-B: Accounting for the Value of New and Free Goods. *American Economic Journal: Macroeconomics*, **17**, 312–344. https://doi.org/10.1257/mac.20210319.

Coyle, D. and Nguyen, D. (2020). *Free Goods and Economic Welfare*. ESCoE Discussion Paper 2020-18, Economic Statistics Centre of Excellence, London.

Data.ai (2024). *State of Mobile India 2024*. San Francisco. Dentsu India (2024). *Digital Advertising in India 2024*. Mumbai.

Feldstein, M. (2017). Underestimating the real growth of GDP, personal income, and productivity. *Journal of Economic Perspectives*, **31**, 145–164.

Goolsbee, A. and Klenow, P. (2006). Valuing consumer products by the time spent using them: An application to the internet. *American Economic Review Papers and Proceedings*, **96**, 108–113.

IAMAI and Kantar (2024). *ICUBE 2024: India's Internet and Digital Report*. New Delhi. IMF (2018). *Measuring the Digital Economy*. Washington, D.C.

IMF (2025). *Balance of Payments and International Investment Position Manual, Seventh Edition (BPM7)*. Washington, D.C.

List, J. and Gallet, C. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, **20**, 241–254.

Murphy, J., Allen, P.G., Stevens, T., and Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, **30**, 313–325.

Nakamura, L. I., Samuels, J., and Soloveichik, R. (2016). Valuing "Free" Media in GDP: An Experimental Approach. *BEA Working Paper WP2016-3*, U.S. Bureau of Economic Analysis.

Nakamura, L. I., Samuels, J., and Soloveichik, R. (2017). Measuring the "Free" Digital Economy within the GDP and Productivity Accounts. *BEA Working Paper WP2017-9*, U.S. Bureau of Economic Analysis.

Nordhaus, W. (2006). Principles of national accounting for nonmarket accounts. In D. Jorgenson, J.S. Landefeld, and W. Nordhaus (eds.), *A New Architecture for the U.S. National Accounts,* University of Chicago Press, 143–160.

NPCI (2024). *UPI Product Statistics 2023–24.* Mumbai.

NSO (2024). *Sources and Methods: National Accounts Statistics.* New Delhi.

NSO/PLFS (2024). *Periodic Labour Force Survey Annual Report, July 2023–June 2024.* New Delhi.

OECD (2020). *A Roadmap Toward a Common Framework for Measuring the Digital Economy.* Paris.

Statistics Canada (2023). Experimental Estimates of Household Digital Intermediation Services. Catalogue No. 13-605-X.

Stiglitz, J., Sen, A., and Fitoussi, J. P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress.* Paris.

Syverson, C. (2017). Challenges to mismeasurement explanations for the US productivity slowdown. *Journal of Economic Perspectives,* **31**, 165–186.

TRAI (2024). *Indian Telecom Services Performance Indicators Report, January–March 2024.* New Delhi.

United Nations, European Commission, IMF, OECD, and World Bank (2025). *System of National Accounts 2025.* United Nations, New York.

Varian, H. (2018). The value of the internet now and in the future. *The Economists' Voice*, **15**, 1–10.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 169-190

Virtual Digital Assets: The Need for Regulatory Framework in India

Praggya Das*

Former Adviser-in-Charge, Monetary Policy Department, Reserve Bank of India

Received: 25 August 2026; Revised: 08 September 2026; Accepted: 09 September 2026

Abstract

The emergence of Virtual Digital Assets (VDAs) has reshaped the global financial landscape, generating a sharp friction between technological innovation and systemic stability. Driven by rapid adoption, the global crypto-asset market capitalisation reached over US$4 trillion by late 2025. Internationally, VDA policies remain deeply fragmented, ranging from an outright ban in China to stringent regulatory oversight in Japan. In India, the VDA market has demonstrated rapid growth, with transaction volumes rising to ₹51,180 crore in 2024-25. This fast expansion presents a critical policy dilemma – balancing the Government's gradual and cautious regulatory approach with the Reserve Bank of India's call to reject private VDAs in favour of Central Bank Digital Currencies (CBDCs). This paper highlights the urgent need for domestic policy coherence and a unified framework for VDAs in India.

*This article is an extended version of a talk delivered at the 28th Annual Conference of the Society of Statistics, Computers and Applications, co-hosted by Savitribai Phule Pune University, on January 5, 2026. Dr. Praggya Das is Former Adviser-in-Charge, Monetary Policy Department, and Former Secretary to the Monetary Policy Committee, Reserve Bank of India, Mumbai. The views expressed are those of the author and usual disclaimer applies.

Key words: Virtual digital asset (VDA); Crypto-assets; Crypto-currencies; Distributed ledger technology; Blockchain; Stablecoins; Central bank digital currency (CBDC).

1. Introduction

Over the last 15 years, there has been an extraordinary rise in the number of new digital instruments, known by various names – cryptocurrencies, digital currencies, crypto-assets, virtual digital assets or virtual assets. These digital assets rely on cryptography to secure transactions and control the creation of additional units. The distributed ledger technology, such as blockchain, is used to construct a ledger (or a database) that is maintained across a network.

Corresponding Author: Praggya Das
Email: praggyadas@gmail.com

The rise of crypto-assets is the result of decades-long search for a cyber solution that enables decentralised peer-to-peer electronic cash network. This evolution was driven by an ideology that views money as a social institution that can be decentralised, effectively bypassing established financial intermediaries. Transfer of digital tokens between unknown counterparts using blockchain technology without a central authority facilitated this movement.

Since the birth of Bitcoin in 2008, the ecosystem has transitioned from a niche experiment to a global phenomenon. The crypto-assets market has experienced explosive growth, rising to a total market capitalisation of more than US$4 trillion by late 2025. However, this rapid expansion has been characterised by unbacked crypto-assets like Bitcoin accounting a large chunk of the market value, despite having no underlying cash flows or intrinsic worth. Central bankers often view this growth much like a Ponzi scheme where value is anchored solely on speculative beliefs rather than fundamental economic utility and underlying cash flows.

While the emergence of crypto and blockchain technology transformed various sectors and opened up fresh avenues for innovation and development, yet the distinct features of digital assets – such as their decentralised structure and dependence on cryptographic security – bring about new risks and challenges. Crypto-assets, since their existence, have displayed significant volatility. Alongside, they have grown in complexity. Virtual Asset Service Providers, which serve as custodians of these assets, are particularly vulnerable to cyberattacks, fraud, and increasing regulatory oversight.

To address the extreme price volatility of traditional cryptocurrencies like Bitcoin, stablecoins entered the crypto arena. Stablecoin is a digital asset the price of which is pegged to stable fiat currencies (like the US$) so users could trade, send money, and store value without losing purchasing power. Stablecoins have commonality and differences with other crypto-assets (which are unbacked) – they are issued on distributed ledgers, typically blockchains like other crypto-assets; and unlike unbacked crypto-assets, they have asset backing and are generally issued and operated in a centralised manner by entities such as crypto firms or financial institutions. A large majority of currently existing stablecoins is denominated in US$.

This article provides an overview of the crypto-assets, their growth in India and the Indian provisions covering them. The global developments to regulate this rapidly developing sector are discussed to show how fragmented the present global regulatory landscape is. The rest of the article is organised in six Sections. Section 2 describes the virtual digital assets ecosystem. The emergence of stablecoins is discussed in Section 3. Section 4 highlights the regulatory fragmentation of global policies governing crypto-assets, focussing both on advanced and emerging market economies; Section 5 focusses on the adoption of VDAs in India and the emergence of regulatory landscape to govern them. Section 6 concludes and discusses the need for policy coherence in India.

2. The virtual digital assets ecosystem

The global crypto-asset or virtual digital asset ecosystem has expanded dramatically and comprises over 10,000 fungible assets with varying characteristics. These are generally categorised into:

- *Unbacked Cryptocurrencies*: Cryptocurrencies are mostly unbacked and most popular type of crypto-assets. They lack both an issuer and any type of underlying asset backing. Users treat unbacked crypto-assets as speculative instruments rather than as a medium of exchange. Their value is determined on supply and demand. These crypto-assets offer limited or no rights for the holder and are usually issued in a decentralised manner. The most popular cryptocurrency is Bitcoin[1].

- *Stablecoins*: Stablecoins are designed to maintain a stable value relative to fiat currencies or commodities and are backed by liquid financial assets. They are issued and operated by centralised entities. Stablecoins are primarily used for crypto trading and their use for cross border payments is expanding. Tether is the most popular stablecoin.

- *Utility and Security Tokens*: Utility tokens are typically offered and sold to raise funds for decentralised platform projects. These tokens are used to access services or represent financial stakes. They are designed to act as a medium of exchange within a platform, for example to pay fees or buy items on a platform. Ether (ETH) and Solana (SOL) are examples of utility tokens. Security tokens, on the other hand, represent ownership on a part of an asset. The value of security tokens is directly related to the valuation of the company or asset issuing the token.

- *Non-Fungible Tokens (NFTs)*: A non-fungible token is a unique digital certificate stored on a blockchain that proves who owns a specific item. Non-fungible means it is one-of-a-kind and cannot be replaced with something identical. Unlike regular money or Bitcoin, where one unit is equal to another, no two NFTs are the same. NFT represent unique digital identifiers for art, in-game items, or real estate that represent value, generated and secured cryptographically and traded electronically, often on a blockchain.

Beyond speculative trading, crypto-assets power useful decentralised software infrastructure. This infrastructure offer several interesting use cases. Some of these include:

- *DeFi (Decentralised Finance)*: DeFi offers an alternative to the existing centralised financial system by using public blockchain networks to conduct transactions without having to rely on centralised service providers such as custodians or central clearing houses. They use smart contracts, which are computer codes stored on public blockchains designed in a way that prohibits intervention and manipulation. There are many advantages from DeFi, but there are also challenges and trade-offs – like it might be harder to enforce investor protection, can be problematic from a privacy perspective, and there may be scalability concerns (Schar, 2022).

- *Tokenisation*: Tokenisation of financial assets and liabilities on programmable digital ledgers is not an improvement of marginal efficiency but a structural shift in financial architecture. It describes how permissioned shared ledgers, programmable financial

[1]In certain regions of Switzerland and for certain payments in UAE, Bitcoin is allowed to function as an alternative point-of-sale medium of exchange. Even for these transactions, at the merchant's end, Bitcoin is instantly converted to local currency and the merchant receives stable fiat, shielding it from Bitcoin's price swings. Fiat currency remains the only official legal tender even in these economies.

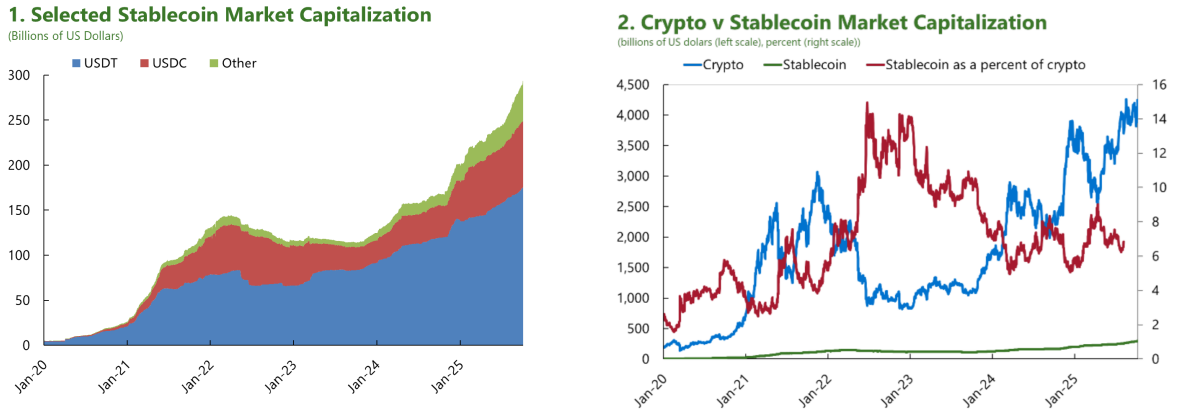
assets, and smart contract-based risk management alter the nature of settlement, liquidity, and systemic risk. However, if public trust is not anchored in digital finance through clear policy frameworks, safe settlements, legal certainty, and international coordination, tokenisation risks may amplify financial instability (Adrian T., 2026).

- *Public Sector Usage*: Countries are using blockchain technology for several projects. These include, inter alia, establishing and maintaining identities for citizens (birth/ death certificates, marriage registration, visas), maintaining interoperable health/ insurance records, land title registry, supply chain management, social security payments, contract and vendor management, voting (enabling digital voting, ensuring eligibility, accurate counting, and auditing), and allowing transparency of government expenditure (OECD, 2018).

Annex 1 provides descriptions of some important terminologies used in the digital assets space.

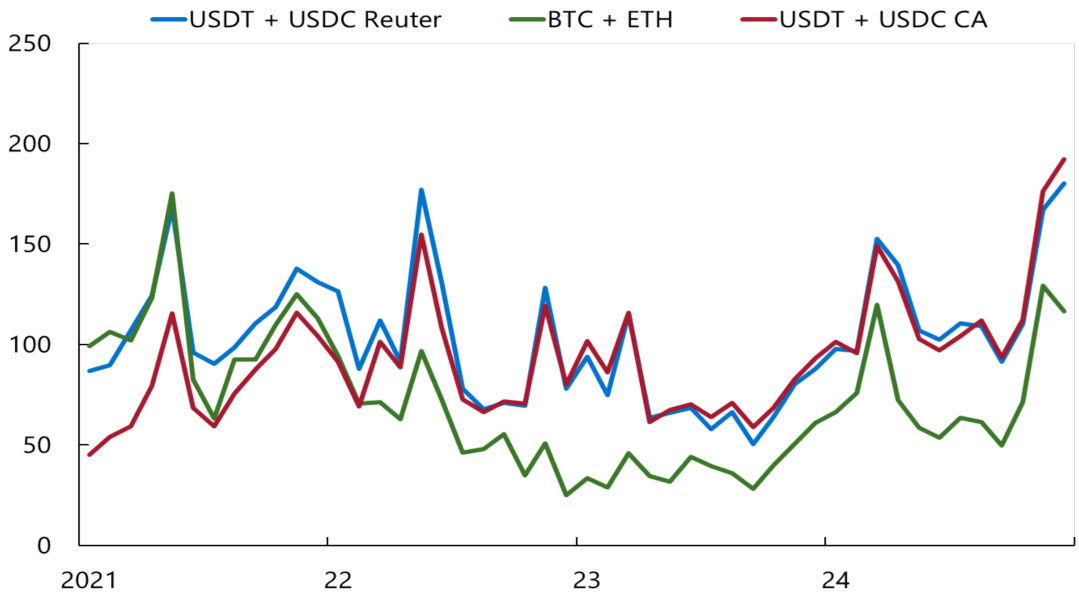
3. The emergence of stablecoins

From among various types of crypto-assets discussed above, the focus, in this Section, is on stablecoins. Stablecoins are crypto-assets that are designed to maintain a steady value by being pegged to another asset. Unlike unbacked cryptocurrencies, stablecoins value is backed by financial reserves, aiming for nominal stability similar to bank deposits. They generally maintain a stable 1:1 value relative to a specified reference asset, typically the US$. Stablecoins represent only a fraction of the wider virtual digital asset capitalisation. The first stablecoins were issued in 2014, but interest in them emerged only a few years later. The stablecoin market scaled with high velocity, doubling in size over a year to approximately US$300 billion by September 2025. USDT (Tether) and USDC (Circle) together command nearly 90 per cent of the market share. While they represent only 7 per cent of total crypto market capitalisation, their role as a bridge between the crypto ecosystem and the real economy is expanding (Figure 1).



Source: Adrian T., *et. al.* (2025)

Figure 1: Stablecoins and crypto market capitalisation



Source: Adrian T., *et. al.* (2025)

Figure 2: Stablecoins and crypto market capitalisation

As per the IMF Paper (2025) stablecoin cross-border flows are already sizable relative to unbacked crypto assets. Though stablecoins account for a small share of total crypto asset market capitalisation, their use in cross-border transactions is substantial (Figure 2).

The primary use cases for stablecoins include:

- Efficient Payments: They enable near-instant, 24x7x365 settlements compared to the traditional banking rails that require 3-5 business days for settlements and are not available round the clock.
- Lower Costs: Remittance costs can be as low as 0.5 per cent with stablecoins, significantly lower than the 2-7 per cent charged by traditional systems.
- On-Chain Settlement: They provide the essential liquid asset needed for the broader growth of asset tokenisation and smart-contract-based transactions.
- Programmable Working Capital and Receivables Financing: Payment networks like Visa supply real-time settlement data to on-chain lenders, allowing business receivables to serve as instant collateral for automated, smart-contract-driven stablecoin loans.

Global usage of stablecoins in absolute terms is led by the Asia, Pacific regions followed by North America. The usage in Africa and Middle East is also significant. The net flows generally emanate from North America to other regions.

'Stability' in stablecoins is maintained through a mechanism of reserve assets held by centralised issuers. The reserve backing of major stablecoins like USDT and USDC

relies on short-term, liquid instruments (Adrian T., et. al. 2025). As of mid-2025, reserve compositions typically included:

- 45-65 per cent of reserves are held in US Treasury Bonds.
- Reverse repurchase agreements, collateralised by US Treasuries account for 10-40 per cent of reserves.
- Cash and Bank Deposits make up 10-15 per cent of reserves.
- Remaining reserves are held in other short-term assets, including commercial paper and financial institution deposits.

Even though stablecoins are designed as stable crypto-assets, the perceived stability of these assets is fragile. The USDC briefly lost its peg in March 2023 following the collapse of Silicon Valley Bank. Such events highlight that stablecoins remain vulnerable to confidence shocks and liquidity risks within their underlying reserves.

Also, unlike traditional dollarisation, which requires physical cash or bank accounts denominated in foreign currency, foreign currency-denominated stablecoins can penetrate an economy rapidly via the internet and smartphones and cause digital dollarisation. Currency substitution facilitated by stablecoin adoption could impinge on monetary sovereignty – a country's ability to exercise full control over its own currency and monetary policy (Adrian T., et. al., 2025).

4. Regulatory fragmentation of policies around crypto-assets

Virtual digital assets (VDAs) have been in existence for the last 15 years, yet public policies to regulate them have gained prominence only recently. Even so, the global regulatory response to crypto-assets is fragmented, as jurisdictions diverge in their approaches for oversight and enforcement.

VDAs are associated with many risks. These include macroeconomic risks, which encompass risks to the effectiveness of monetary policy, capital flow volatility, and fiscal risks. There are also serious concerns about financial stability, financial integrity, legal risks, consumer protection, market integrity, and contestability. Some risks are inherent to the technology underpinning crypto-assets, while others stem from the lack of policies or their enforcement (IMF Policy Paper, 2023). Since crypto-assets have implications for both macroeconomic and financial stability, a comprehensive regulatory and supervisory oversight of crypto-assets should be a baseline to address these risks.

4.1. International frameworks for crypto-assets

For the crypto-assets sector, multilateral bodies like the International Monetary Fund (IMF), the Financial Stability Board (FSB) have developed a global framework of recommendations and standards. At the request of the Indian G20 Presidency, the IMF and the FSB have come out with a paper to synthesise their and standard setting bodies' (SSBs') policy recommendations and standards (Financial Stability Board, 2023).

The IMF and the FSB have also set out a roadmap to ensure effective, flexible, and coordinated implementation of the comprehensive policy response for crypto-assets. As per their recommendations, to address risks to financial integrity and mitigate criminal and terrorist misuse of the crypto-assets sector, jurisdictions should implement the Financial Action Task Force (FATF) anti-money laundering and counter-terrorist financing (AML/CFT) standards that apply to VDAs and VDA Service Providers (VDASPs).

Separately, the Basel Committee on Banking Supervision (BCBS) of the Bank for International Settlements (BIS) has come out with guidelines for banks in respect of their exposures to crypto-assets. Crypto-assets are defined as private digital assets, and which can be used for payment or investment purposes or to access a good or service. All on- or off-balance sheet amounts that give rise to credit, market, operational and/or liquidity risks are included in banks' exposures. The Basel Framework has introduced prudential standards for crypto-assets for banks (Bank for International Settlements, 2026).

4.2. Global regulatory approaches

Policymakers around the world have been developing a variety of national approaches to manage crypto-assets. These vary significantly from outright ban to targeted restrictions to combinations of regulation-supervision-taxation to the other extreme of granting unbacked tokens legal tender status. We discuss here the policy direction of some of the major economies.

Advanced Economies:

- **European Union**: The framework to regulate crypto-assets in EU called – Markets in Crypto-Assets (MiCA) regulation – was enacted in June 2023. EU has a fully regulated uniform legal framework across all 27 member states. The key provisions under MiCA for those issuing and trading crypto-assets (including asset-reference tokens and e-money tokens) cover transparency, disclosure, authorisation and supervision of transactions. The new legal framework allows EU-wide passporting for licensed Crypto-Asset Service Providers. Strict capital reserve requirements are prescribed for stablecoins. The Act supports market integrity and financial stability by regulating public offers of crypto-assets and by ensuring consumers are better informed about their associated risks.[2]

- **Japan**: Japan was one of the early adopters of a dedicated digital asset statutory framework. Japan's crypto laws have been updated over years through legislative amendments, the latest landmark reform being passed in July 2026. Cryptocurrency is a legal financial instrument in Japan. The country strictly regulates crypto exchanges to protect regular users from hacks and fraud, making it one of the safest yet most tightly controlled crypto markets in the world. Crypto-Asset Exchange Service Providers need to be registered to function in Japan. Crypto-assets functioning as payments, such as stablecoins, are regulated under the Payment Services Act (PSA), overseen by the Financial Services Agency (FSA). PSA strictly restricts stablecoin issuance to licensed banks, registered money transfer agents, and trust companies.

[2]https://www.esma.europa.eu/esmas-activities/digital-finance-and-innovation/markets-crypto-assets-regulation-mica

- **Singapore**: Crypto-assets are regulated in Singapore by the Monetary Authority of Singapore (MAS). There are two Acts that govern crypto-assets – the Payment Services Act (for registering crypto-asset players); and the Financial Services and Markets Act (for MAS oversight to Singapore-incorporated entities that provide digital token services exclusively to overseas customers). Regulations for stablecoins in Singapore allow single-currency stablecoins pegged to the Singapore Dollar or G10 currencies, mandating strict reserve backing, high-liquidity assets, and rapid par-value redemptions.

- **Taiwan**: The Virtual Asset Service Act was passed by the legislature in Taiwan on June 30, 2026. The Act outlines licensing requirements for virtual asset service providers to obtain approval from the Financial Supervisory Commission (FSC) to operate. This recent development is a decisive shift from anti-money laundering registration alone that was in force until recently. For stablecoins, the new law requires full reserve assets to be held in financial institutions within Taiwan. The assets are subject to regulatory audits. Algorithmic stablecoins, which are not backed by assets, are not permitted in Taiwan.

- **United Kingdom**: The crypto-asset regulation regime in UK is underpinned by the Financial Services and Markets Act 2000 (Crypto-assets) Regulations 2026 (the Crypto-assets Regulations), which were passed by Parliament in February 2026. These Crypto-assets Regulations brought a broad range of crypto-asset activities within the Financial Conduct Authority's (FCA's) regulatory perimeter for the first time, moving well beyond the anti-money laundering and financial promotions standards that previously defined crypto-assets' role in this market.[3] Under the new Act, UK has come out with a Crypto-asset Roadmap and a comprehensive regime that sets standards, strengthens consumer protections, and positions UK as a trusted, competitive home for responsible crypto-asset innovation.

- **United States**: The US crypto laws are governed by a multi-agency framework and are evolving. The core regulatory bodies are (i) Securities and Exchange Commission (SEC) for digital assets classified as securities; (ii) Commodity Futures Trading Commission (CFTC) to regulate digital assets recognised as commodities, such as Bitcoin; and (iii) Financial Crimes Enforcement Network (FinCEN) that enforces Anti-Money Laundering (AML) and Know Your Customer (KYC) rules. Key laws/regulations that govern crypto-assets are: (a) GENIUS (Guiding and Establishing National Innovation for US Stablecoins) Act for stablecoins that became law in July 2025. The Act requires a 1:1 reserve backing, monthly audits, and licensing requirements for US$-pegged stablecoins; (b) the proposed 'Regulation Crypto-Assets (RCA)' would provide crypto asset entrepreneurs with clear pathways to raise capital under the federal securities laws; and (c) the proposed CLARITY Act to clearly divide jurisdiction between the SEC and CFTC and define digital commodity structures. [4]

Emerging Market Economies (EMEs):

[3] https://www.fca.org.uk/publications/policy-statements/cryptoasset-regime

[4] https://www.sec.gov/newsroom/speeches-statements/atkins-statement-regulation-crypto-assets-081826

The rapid growth of crypto-asset sector presents complex macroeconomic and regulatory challenges for EMEs, where citizens increasingly adopt these assets for savings and investment, or for hedging against volatile domestic currencies. This rapid adoption of crypto-assets complicates regulatory oversight, exposing EMEs to multifaceted risks. High price volatility, alongside opaque and concentrated trading structures, drive acute market and liquidity risks for investors. Meanwhile, excessive leverage in DeFi protocols and weak market governance introduce significant credit risks, which is compounded by operational vulnerabilities to cyberattacks. Globally, widespread crypto integration threatens currency substitution, accelerates bank disintermediation as deposits shift away, and complicates capital flow management via untraceable cross-border payments (BIS, 2023). Select EME country policy response to the rapid crypto-asset growth is discussed below.

- **Brazil**: Brazil regulates its digital assets market (including that for stablecoins) under the Brazilian Virtual Assets Law, which establishes a comprehensive framework overseen by its central bank, Banco Central do Brasil (BCB) for virtual asset service providers. For crypto-assets functioning as securities, the legal jurisdictions is with Securities Commission (CVM). BCB Resolutions that came into force from February 2026 mandates formal licensing for local operators, restrict cross-border offerings from unapproved foreign entities, enforce strict asset segregation, and phase in the Crypto Travel Rule for transaction transparency.[5]
- **China**: China maintains a comprehensive ban on all crypto transactions, offshore exchange services targeting domestic residents, token issuance, and bitcoin mining. The government prioritises the adoption of its central bank digital currency, the e-CNY. Legal tender status is reserved strictly for fiat and the digital yuan (e-CNY).
- **India**: Cryptocurrency is legal to buy, sell, and hold in India, but it is not recognised as legal tender. India does not have a single, comprehensive law governing crypto, but it enforces a strict tax and anti-money laundering framework. India case is discussed in detail in Section 5.
- **Indonesia**: Cryptocurrency is legal to trade and own as a digital financial asset in Indonesia. Bank Indonesia strictly enforces the rupiah as the only legal tender for transactions, making crypto payments illegal. Regulatory oversight of the sector is carried out by the Financial Services Authority (locally called OJK – Otoritas Jasa Keuangan), with effect from January 2025. All crypto market operators – exchanges, clearing institutions, custodians, and digital asset traders – need to secure license from OJK. No trading is allowed in crypto assets that fall outside the OJK approved list. Stablecoins are treated strictly as transactional instruments – it can be bought, sold, tokenised on licensed platforms, but cannot be used to pay for goods or services. OJK is preparing to bring crypto assets deeper into the country's financial system.[6]
- **Russia**: Russia's comprehensive cryptocurrency legislation formally took effect on September 1, 2026. The framework grants digital currencies legal status as property while placing them under the strict oversight of the Bank of Russia.[7] Trading and

[5]https://www.bcb.gov.br/en/pressdetail/2639/nota

[6]https://indonesiacrypto.network/blog/ojk-prepares-to-integrate-crypto-into-indonesias-financial-system

[7]https://www.cbr.ru/eng/press/event/?id=32724

custody is permitted through licensed domestic intermediaries. Russia explicitly permits the use of liquid cryptocurrencies like Bitcoin for cross-border trade settlements. However, the legislation maintains a strict ban on cryptocurrencies for any domestic payments for goods, or services within the country.

- **South Africa**: In South Africa, cryptocurrencies are legally classified as financial products rather than official legal tender. According to the Financial Sector Conduct Authority (FSCA), all Crypto Asset Service Providers (CASPs) must be formally licensed to operate.[8] Furthermore, under the Financial Intelligence Centre Act (FICA), digital asset platforms are accountable institutions required to implement strict identity verification and anti-money laundering controls. Stablecoins are legally classified as financial products and only CASP licensed entities can deal with them. The South African Reserve Bank (SARB) actively regulates stablecoins, focusing on monetary stability, financial surveillance, and exchange controls. Stablecoins are not recognised as legal tender and are restricted for use in domestic payments.

4.3. The challenge of policy fragmentation

The crypto-assets regulatory framework of different countries is at different level of development. Without proper regulations, crypto markets become unsafe and highly vulnerable to frauds with regulatory deficit empowering illicit actors and compromising consumer safety. On the other hand, stringent supervisory frameworks effectively mitigate systemic exploitation and drive away non-compliant entities toward weaker regulatory regimes. Fragmented frameworks also hinder international cooperation, making cross-border supervision and crisis management increasingly difficult.

4.4. Data reporting on virtual digital assets

Given that the size of virtual digital assets is large and there are global linkages, IMF has come out with compilation guidance on crypto-assets.

IMF suggests that compilers may select most suitable data sources to produce consistent and reliable statistics on crypto-assets, since the nature of crypto related activities is complex. The source of information could be administrative data (regulatory filings, tax records, international transaction data, and reports on crypto-asset losses, thefts, and seizures); financial reports (of mining companies, stablecoin issuers, crypto exchanges, and publicly listed corporations), or alternative (commercial) data sources.[9]

In the absence of dedicated surveys or established statistical reporting systems, regulatory reporting offers a practical starting point for compiling crypto-asset data. These reports may, inter alia, help in investor protection, market oversight, and compliance enforcement.

[8] https://www.fsca.co.za/New-Financial-Service-Provider/

[9] https://unstats.un.org/unsd/nationalaccount/aeg/2025/M28/M28_9_Crypto_Assets_Pres.PDF

5. Adoption of virtual digital assets in India and the regulatory landscape

The virtual digital asset industry in India is fast growing. India's cryptocurrency adoption has evolved rapidly from early grassroots trading into a large retail-driven ecosystem.

5.1. The size of Indian crypto space

The Indian VDA industry is home to over 1,250 startups. These ventures have raised US$3.5 billion in cumulative funding since 2020. Indian Web3 companies have seen a steep upward trajectory of growth and accounts for 15 per cent of the global Web3 developer talent. The base is young, led by students, of which, about 90 per cent are male. India has become the top blockchain adoption market and a leading hub for founders and developers driving innovation across Web3 (India Web3 Landscape Report, 2025).

As per TRM Labs, a leading blockchain intelligence and data analytics company, India has emerged as a global frontrunner in the adoption of VDA. In 2025, India ranked number one globally in the Global Crypto Adoption Index when adjusted for the size of the economy.[10] TRM analysis suggests that India's position at the top is likely due to its large, young population's rising interest in digital assets, supported by a crypto-fluent middle class and a thriving developer ecosystem. Interest in crypto as an asset class is also growing among institutional and high-net-worth investors, reinforcing India's accelerating crypto adoption.

According to news reports[11], India's crypto market jumped 41 per cent year-on-year in 2024-25, reflecting growing use of virtual digital assets amid evolving government policy environment. As reported by The Hindu, data presented in the Parliament reveals the Ministry of Finance collected ₹511.8 crore as Tax Deducted at Source (TDS) from cryptocurrency trades. At a 1 per cent TDS rate, this equates to ₹51,180 crore (UD$6.14 billion) in total transactions during 2024-25. India's TDS on crypto transfers, was designed to track transactions and ensure proper reporting of gains. The rule has become a key source of data on market size. India's cryptocurrency market has recorded steady growth over the past three years, with transactions rising from ₹22,130 crore in 2022-23 to ₹36,270 crore in 2023-24, and further to ₹51,180 crore in 2024-25.

Crypto adoption in India was initially a metro phenomenon. Of late, crypto adoption is moving from being metro centric to Tier 2, 3 and 4 cities driven. As per CoinSwitch, a leading VDA service provider in India, these cities have a share of over 75 per cent of their user base. Uttar Pradesh leads with 13 per cent of total invested value, ahead of traditional centres like Maharashtra and Karnataka. Crypto investor base is young with 18-25 year investors holding a quarter of investments and 26-35 year group holding about a half of the overall crypto investments. This reflects a massive interest in blockchain and Web3 technology among the younger generation (CoinSwitch, 2025). Details on unfolding of India's crypto story are also available from reports by other players (for example, CoinDCX, 2025).

[10] https://www.trmlabs.com/reports-and-whitepapers/2025-crypto-adoption-and-stablecoin-usage-report

[11] https://aibc.world/news/india-posts-strong-crypto-growth/ and https://www.thehindu.com/business/Economy/crypto-transactions-in-india-in-2024-25-tax-collection-data-shows/article70372854.ece

Bharat Web3 Association (BWA) is a leading VDA industry body, that has been pivotal in bridging the gap between innovation and regulation. BWA comprises of over 40 members including major crypto players like CoinSwitch, CoinDCX, Coinbase, Mudrex, Delta, and WazirX. BWA is working towards establishing industry standards, cybersecurity guidelines, and fostering stakeholder dialogue to build a compliant VDA ecosystem in the country. BWA collaborated extensively with FIU-IND to develop Red Flag Indicators (RFIs) for detecting suspicious transactions, ensuring the VDA sector aligns with global financial integrity standards. In absence of formal regulation, there are ongoing proposals for self-regulatory oversight.

5.2. VDA's regulatory landscape in India

India's approach for VDA regulation has been cautious with the aim to protect users, prevent fraud, and curb money laundering, while allowing innovation to continue. It operates under a stringent tax and anti-money laundering framework without dedicated recognition. While buying and holding VDAs is legal, transactions face a flat 30 per cent rate of income tax on gains. The VDA sector and its regulatory journey in India has seen its ups and downs.

5.2.1. The RBI's caution in 2013

In December 2013, the Reserve Bank of India (RBI) cautioned the users, holders and traders of virtual currencies (including Bitcoins), about the potential financial, operational, legal, customer protection and security related risks that they are getting exposed to.[12] The RBI mentioned that creation, trading or usage of VCs including Bitcoins, as a medium for payment are not authorised by any central bank or monetary authority. No regulatory approvals, registration or authorisation has been obtained by the entities concerned for carrying on such activities. As such, they may pose several risks to their users and the users were expected to exercise caution.

In February 2017, the RBI issued another caution and informed the users that it had not authorised any entity to operate or deal with Bitcoin or any virtual currency. Any user, holder, investor, trader, *etc.* dealing with virtual currencies will be doing so at their own risk. Later in December 2017, in the wake of significant spurt in the valuation of many virtual currency and rapid growth in initial coin offerings, the RBI reiterated its concern.

5.2.2. The banking ban in 2018 and the legal reversal

In April 2018, the RBI barred all regulated financial institutions and banks from servicing entities dealing in or exchanging cryptocurrencies.[13] The ban applied to maintaining accounts, registering, trading, settling, clearing, giving loans against virtual tokens, accepting them as collateral, opening accounts of exchanges dealing with them and transfer/receipt of money in accounts relating to purchase/sale of virtual currencies. The regulated entities which were providing such services were advised to exit the relationship within three months.

The Supreme Court of India, however, set aside the RBI's 2018 circular, in March

[12] https://www.rbi.org.in/scripts/BS_PressReleaseDisplay.aspx?prid=30247
[13] https://www.rbi.org.in/scripts/NotificationUser.aspx?Id=11243&Mode=0

2020 as the RBI could not show empirical evidence of actual harm to regulated banks.[14] The lifting of ban in March 2020, allowed startups like CoinSwitch and CoinDCX to resume formal banking partnerships and spark explosive pent-up retail growth.

In May 2021, the RBI advised its regulated entities that for dealing in virtual currencies, they may continue to carry out customer due diligence processes, in line regulations under KYC, AML, Combating of Financing of Terrorism, Prevention of Money Laundering Act (PMLA), and Foreign Exchange Management Act (FEMA) for overseas remittances.

5.2.3. Formal definition of VDA and taxation in 2022

In India, virtual digital assets are formally defined under Section 2(47A) of the Income Tax Act, 1961 as any information, code, number, or token generated through cryptographic means or otherwise, which provides a digital representation of value, can be stored, transferred, or traded electronically, and includes Non-Fungible Tokens (NFTs). In line with the global definition, VDAs in India also include cryptocurrencies, tokens, and NFTs.

The current taxation related aspects for VDAs and its reporting requirement are covered in the Income Tax Act 1961 as amended in 2022. Income arising from dealing in VDAs is taxed at a flat 30 per cent rate, with no set-off for losses other than cost of acquisition. TDS of 1 per cent is also applied when payment is made to a resident for transfer of VDA. The Act also requires that specified reporting entities must report transactions involving crypto-assets in a prescribed form and timeline (Annex 2).

5.2.4. Advertising guidelines in 2022

In February 2022, the Advertising Standard Council of India (ASCI), an independent, voluntary self-regulatory organisation, brought out the requirements that all advertisements for VDA products and VDA exchanges must follow.[15] The guidelines are detailed and cover, inter alia, the following aspects:

- All advertisements must carry the disclaimer "Crypto products and NFTs are unregulated and can be highly risky. There may be no regulatory recourse for any loss from such transactions." Such a disclaimer must be PROMINENT and UNMISSABLE by an average consumer.
- No advertisement for VDA products or exchanges may show a minor directly dealing or talking about the product.
- VDA products may not be compared to any other asset class which is regulated.
- Since this is a risky category, celebrities or prominent personalities who appear in VDA advertisements must take special care to ensure that they have done their due diligence, so as not to mislead consumers.

[14] https://indiankanoon.org/doc/12397485/

[15] https://www.ascionline.in/the-asci-code-guidelines/

5.2.5. Registration requirement under FIU-IND in 2023

The Financial Intelligence Unit – India (FIU-IND) in March 2023, brought out guidelines for registration of VDA Service Providers. The VDASPs are now mandated, inter alia, to:

- Register with FIU-IND,
- Comply with the various provisions of PMLA,
- Establish process for customer due diligence,
- Maintain records,
- Set up internal control procedures,
- Ensure training of employees, and
- File suspicious transaction report with FID-IND.

5.2.6. Other legislative developments

In February 2023, government enhanced its efforts against digital fraud by integrating reporting for VDA-related frauds into the National Cyber Crime Reporting Portal (NCRP) and its associated helpline, 1930, for consumer grievance redressal. [16]

In December 2023 FIU-IND issued show-cause notices to nine major off-shore VDA exchanges for non-compliance with VDA regulations. The access to non-compliant platforms was blocked. By 2025, some of the foreign exchanges came back and registered with FIU-IND, reflecting broader industry shift towards regulatory compliance (CoinSwitch, 2025).

Union Budget 2025 introduced new Section 285BAA in the Income-Tax Act, mandating crypto-asset reporting by entities (like exchanges) for transactions from April 1, 2026, to combat tax evasion and enhance transparency.[17]

In their March 2026 topical report, "Crypto Assets in India: Assessing the Case for Regulation," the Gujarat National Law University underscores that the need for tailored, forward-looking crypto regulation in India is both urgent and undeniable. The report examines various regulatory approaches with SEBI (Securities and Exchange Board of India) as primary regulator, with RBI as primary regulator, with a Ministry of Electronics and Information Technology (MeitY)-led oversight, a self-regulatory organisation model, a new dedicated regulator, and a multi-agency (hybrid) model. The report looks at international instances of collaborative regulation and maps out how India's regulatory bodies could coordinate to cover all facets of VDA activity.

A recent Report of the Standing Committee on Finance in July 2026 recommends that "the Government comprehensively examine the need for an appropriate statutory and regulatory framework for Virtual Digital Assets. Pending the establishment of a comprehensive legislative framework, the Government may consider introducing an interim mechanism through recognised Self-Regulatory Organisations (SROs) operating under the oversight of the designated regulator. Such a framework should prescribe minimum standards of governance, transparency, disclosure, investor protection, grievance redressal, compliance with

[16] `https://cybercrime.gov.in/Webform/CrimeCatDes.aspx`
[17] `https://www.incometaxindia.gov.in/w/section-285baa`

prescribed codes of conduct and appropriate regulatory oversight, so as to mitigate risks arising from the existing regulatory vacuum while promoting market discipline and safeguarding investor interests" (Standing Committee on Finance, 2026).

5.2.7. International commitments by India

India has officially subscribed to two primary international reporting frameworks to curb tax evasion and illicit financial flows related to VDAs: the Financial Action Task Force (FATF) Standards[18] and the OECD's Crypto-Asset Reporting Framework (CARF).

The FATF routinely monitors money laundering, terrorist financing, and proliferation financing (*i.e.*, the financing of weapons of mass destruction) across jurisdictions to protect global financial system from illicit activity. In recent years it has expanded its focus to VDAs and VDASPs.

For implementation of CARF, India's Presidency of the G20 in 2023 lent decisive momentum and India actively contributed to the development of these standards through its participation in global forums. The obligation to report transactions in crypto-assets is now provided for under section 509 of the Income-tax Act, 2025. The Central Board of Direct Taxes (CBDT), Ministry of Finance has prepared a Guidance Note to assist Reporting Crypto-Asset Service Providers in understanding and discharging their obligations under CARF (Government of India, 2026).

Thus, India has fully integrated these international frameworks into domestic law. While there is no single overarching law for VDAs in India, multi layered regulatory architecture across Income Tax Act, PMLA and international reporting framework legally govern VDAs.

5.3. The use of blockchain technology for CBDC

While private sector growth in the virtual digital assets was continuing, the Government was working towards utilising the underlying technology in public sector. To strengthen the governance through blockchain technology, a National Blockchain Framework was launched in September 2024. In the public sector, the blockchain technology is being used in several areas – for storage and retrieval of certificates and documents, for logistics to track movement of goods, for maintaining judiciary data, for property transactions, in telecom ecosystem, for digital rupee by RBI, for debenture covenant monitoring by NSDL (National Securities Depository Limited), *etc.*[19]

Regarding digital rupee, using Distributed Ledger Technology (DLT) and token-based blockchain architecture, the RBI has brought out Central Bank Digital Currency (CBDC). A digital form of India's physical currency, CBDC or digital Rupee or e₹, is a legal tender issued by the RBI. It is exchangeable par-to-par with physical cash, is backed directly by the central bank, and is stored in mobile digital wallets. It can be used to receive/send money, and make payment for transactions, just like any physical currency.

The RBI says "CBDCs are digital tokens like stablecoins yet they are inherently

[18] https://www.enforcementdirectorate.gov.in/international-cooperation/fatf/

[19] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2182023&reg=3&lang=2

superior since they satisfy all the attributes that money should have – fiat, single, trusted and representing value – and do not pose many of the risks associated with stablecoins" (Sankar, 2025).

CBDC is being used for both retail and wholesale transactions. In the retail space, wallet-to-wallet transfers were introduced first. Interoperability with UPI (Unified Payments Interface) also facilitated users to pay to merchants using their e₹wallets. Feature of discovering existing CBDC wallets with Bharat Interface for Money (BHIM) app has further facilitated CBDC payments. Programmability feature of CBDC (embedding rules, logic and conditions) allows programmable money (CBDC) transfers by individuals. In the wholesale space, government welfare and subsidy payments are being done using CBDC.

RBI is also working on new use cases that include, inter alia, pilots in various government agencies' welfare payments schemes leveraging CBDC's programmability feature to ensure productive utilisation of public funds; offline CBDC, both proximity based (Near Field Communication) and non-proximity based (SMS); secondary market trades settlement of G-Secs and inter-bank borrowing in CBDC; and providing a framework for testing of innovative products/services leveraging CBDC under the CBDC and Asset Tokenisation (CAT) sandbox. Most importantly, the RBI is actively engaging in developing cross-border CBDC arrangements with other countries on a bilateral and multilateral basis (RBI, 2026). Under India's Chairmanship of BRICS 2026, the BRICS Summit is focusing on official state-backed CBDCs and blockchain-based government settlement tokens. The RBI is working on connecting official CBDCs and respective fast payment systems of BRICS members to streamline cross-border trade, offer payment choices and lower transaction costs.[20]

6. Conclusion and policy path going forward

Virtual digital assets (VDAs) offer transformative benefits – ranging from streamlined payments and asset tokenisation to enhanced transaction transparency. These immense potential of VDAs is heavily caveated by structural vulnerabilities. Without a centralised statutory regulator, existing compliance frameworks cannot fully shield users from rampant scams, systemic volatility, and sophisticated regulatory arbitrage.

India stands at a decisive crossroad. On the one hand , domestic adoption and market interest in VDAs continue to grow rapidly. On the other hand, policy approach for VDAs remains fragmented – there is a marked divergence between the views of the Government and the RBI.

The RBI remains highly cautious about VDAs, particularly stablecoins. The concern is rooted in safeguarding monetary authority, financial stability, and macroeconomic integrity. Since VDAs operate on decentralised rails and bypass formal financial controls, their widespread adoption threatens to cause digital dollarisation of the economy, which would strip the central bank of its ability to control money supply, manage inflation, and implement monetary policy. For stablecoins specifically, the risk of disintermediating the banking system by replacing low-cost bank deposits, threatens driving up credit costs and weakening the overall financial system. Private stablecoins also present severe capital flow

[20] https://www.reuters.com/world/india/brics-nations-discuss-linking-payment-systems-cbdcs-rbi-chief-says-2026-08-11/

management challenges, divert sovereign seigniorage revenue to private operators, often outside the home jurisdiction, and risk facilitating illicit activities by bypassing traditional KYC and AML/CFT regulations. Given India's highly efficient public payments landscape through UPI and proactive promotion of its own CBDC as a safer, fully anchored sovereign alternative for digital innovation, the RBI maintains that the severe systemic risks of private cryptocurrencies far outweigh their purported benefits, and "they do not serve a purpose that cannot be served better by fiat money" (Sankar, 2022 and 2025).

Given that global growth in crypto-assets, both globally and domestically, is rapidly rising, an outright rejection of VDAs exposes India to more risks than a strategically regulated environment. Critique of the central bank's rigid preference for state control points out that ignoring the operational benefits of stablecoins – such as 24x7 atomic settlement, smart-contract programmability, and seamless integration across digital markets – is analytically unconvincing. If tokenised dollars get embedded in global trade and cross border settlements, there is possibility of digital dollarisation, capital flow leakage outside domestic control and partial bypassing of local payment systems (Gurumurthy, 2025). Finance Minister in a speech has signalled openness to engagement with the digital assets sector, emphasising that nations cannot insulate themselves from innovations like digital currencies and stablecoins. The Government's focus is on innovation, competitiveness, and strategically adapting to new monetary architectures without risking global exclusion (Sitharaman, 2025).

The Government is gradually bringing in multi-layered regulations for the VDA sector across Income Tax Act, the PMLA, mandatory registrations of VDA Service Providers, and for international reporting. Yet, recent intelligence highlighting how residents utilise private wallets and offshore fintech platforms to convert VDAs into domestic purchasing power – closed-loop gift vouchers for gold, fuel, and even groceries – exposes a critical policy gap.[21] This mechanism demonstrates that market participants actively exploit regulatory arbitrage through non-crypto, retail intermediaries, bypassing the foreign exchange frameworks and tax surveillance without interacting with legally registered domestic VDASPs.

Consequently, an effective VDA regulatory framework cannot focus strictly on the blockchain layer or exchange-level compliance. Achieving true policy coherence requires an approach that bridges VDA regulations with traditional payments and retail oversight. Policymakers must enforce stringent reporting, compliance, and audit requirements on domestic fiat off-ramps – including bulk gift voucher aggregators, Prepaid Payment Instrument (PPI) issuers, and cross-border settlement channels. Furthermore, as data reporting mechanisms mature, establishing regulatory reporting lines will prove essential for tracking cross-border crypto flows, investor activity, and peer-to-peer leaks outside regulated intermediaries.

Finally, pending the enactment of a comprehensive legislative framework, the Government may consider the Standing Committee on Finance recommendation to introduce an interim mechanism through recognised Self-Regulatory Organisations (SROs). Operating under the designated regulatory oversight, SROs can establish minimum standards of governance, transparency, disclosure, investor protection, grievance redressal and compliance with prescribed codes of conduct. This would instil market discipline and safeguard investor

[21] https://m.economictimes.com/news/india/indians-are-using-crypto-to-buy-groceries-fuel-and-gold-through-overseas-gift-card-platforms/amp_articleshow/133895144.cms

interests and mitigate risks arising from the absence of an overarching regulatory framework for the VDA sector.

India's path toward *Viksit Bharat @ 2047* requires aligning the gap between the Government's innovation agenda and the RBI's stability imperatives. By pairing targeted VDA rules with robust surveillance over non-crypto payment endpoints, India can build a coherent regulatory framework that preserves national financial sovereignty while positioning the country at the forefront of digital financial innovation.

Conflict of interest

The author does not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Tobias, A. (2026). *Tokenized Finance*. International Monetary Fund Notes 2026/001, April 2026. https://www.imf.org/-/media/files/publications/imf-notes/2026/english/insea2026001.pdf.

Adrian, T., Bains, B., Bechara, M., Cerutti, E. *et al.* (2025). *Understanding Stablecoins*. IMF Departmental Paper, December 4, 2025. https://www.imf.org/en/publications/departmental-papers/issues/2025/12/02/understanding-stablecoins-570602.

Bank for International Settlements (2023). *Financial Stability Risks from Cryptoassets in Emerging Market Economies*. Consultative Group of Directors of Financial Stability (CGDFS), BIS Papers No 138, August 2023. https://www.bis.org/publications/paper-138-financial-stability-risks-cryptoassets-emerging-market-economies.pdf.

Bank for International Settlements (2026). *Scope and Definitions Crypto-asset Exposures*. January 1, 2026. https://www.bis.org/committees/bcbs/basel-framework/basel-framework.pdf

CoinDCX (2025). *H1 2025 Report: Unfolding India's Crypto Story*. August 5, 2025. https://coindcx.com/blog/wp-content/uploads/2025/08/CoinDCX-Half-yearly-Report-25.pdf.

CoinSwitch (2025). *Virtual Digital Assets Decoded, Understanding Technology, Legality and Frauds*. https://coinswitch.co/switch/wp-content/uploads/2025/12/Handbook-VDA-Decoded-PrintFile-New.pdf.

Financial Stability Board (2023). *IMF-FSB Synthesis Paper: Policies for Crypto-Assets*. September 7, 2023. https://www.fsb.org/2023/09/imf-fsb-synthesis-paper-policies-for-crypto-assets/.

Gujarat National Law University (2026). Report on Crypto-assets In India: Assessing the Case for Regulation. March 2026. https://gnlu.ac.in//Content/gnlu/pdf/odrd/Abridged%20Version_Report_Crypto%20Assets%20Project.pdf.

Gurumurthy, R. (2025). *Stablecoins, CBDC and RBI's Selective Logic BasisPoint*. December 21, 2025. https://basispointinsight.com/Story/stablecoins--cbdcs--and-the-rbi-s-selective-logic_bdab527c783a.html.

Government of India (2026). *Crypto-asset Reporting Obligations – Guidance Note*. Department of Revenue, Central Board of Direct Taxes, Foreign Tax and Tax Research Division, July 24, 2026. https://www.incometaxindia.gov.in/documents/d/guest/e-book-crs-and-fatca-guidance-note-faqs-pdf.

IMF Policy Paper (2023). *Elements of Effective Policy for Crypto-assets*. February 2023. https://www.imf.org/-/media/files/publications/pp/2023/english/ppea2023004.pdf.

India Web3 Landscape Report 2024. *Hashed Emergent*. https://bwaind.in/wp-content/uploads/2025/03/India-Web3-Landscape-Report-2024.pdf.

India Web3 Landscape Report 2025. *Hashed Emergent*.

Organisation for Economic Co-operation and Development (OECD, 2018). *Blockchain and its Use in the Public Sector*. Observatory of Public Sector Innovation, June 20, 2018. https://www.slideshare.net/slideshow/blockchain-and-its-use-in-the-public-sector-oecd/103148771.

Reserve Bank of India (2026). Annual Report 2025-26, May 29, 2026. https://www.rbi.org.in/Scripts/AnnualReportPublications.aspx?year=2026.

Sankar, T. R. (2022). *Cryptocurrencies – An Assessment*. Keynote address at the Indian Banks Association 17th Annual Banking Technology Conference and Awards. https://www.rbi.org.in/Scripts/BS_SpeechesView.aspx?Id=1196.

Sankar, T. R. (2025). *Stablecoins – Do They Have a Role in the Financial System*. Keynote address at the Mint Annual BFSI Conclave, December 12, 2025. https://www.rbi.org.in/Scripts/BS_SpeechesView.aspx?Id=1538.

Schar, F. (2022). *DeFi's Promise and Pitfalls*. Finance and Development, September 2022. https://www.imf.org/-/media/files/publications/fandd/article/2022/september/schar.pdf.

Sitharaman, N. (2025). *Seeking Prosperity in Turbulent Times*. Inaugural Address by Smt. Nirmala at the Kautilya Economic Conclave, October 3, 2025. https://iegindia.org/inaugural-address-by-smt-nirmala-sitharaman-at-the-kautilya-economic-conclave-2025/.

Standing Committee on Finance (2026). *Thirty-Sixth Report (Eighteenth Lok Sabha) on the Bill 'The Securities Markets Code 2025 of the Ministry of Finance (Department of Economic Affairs), Government of India*. July 2026. https://www.medianama.com/wp-content/uploads/2026/07/app_lsscommittee_Finance_18_Finance_36.pdf.

ANNEXURE

Annex-1: Glossary of Select Terms from the VDA Space

Sr.	Term	Description
1	Blockchain	A 'blockchain' is a particular type of data structure used in some distributed ledgers which stores and transmits data in packages called "blocks" that are connected to each other in a digital 'chain'. Blockchains employ cryptographic and algorithmic methods to record and synchronize data across a network in an immutable manner.
2	Cryptocurrencies	Cryptocurrencies are digital currencies that rely on cryptographic techniques to achieve consensus. *Example*: Bitcoin (BTC) and Ether (ETH).
3	Distributed ledger technology (DLT)	DLT refers to the protocols and supporting infrastructure that allow computers in different locations to propose and validate transactions and update records in a synchronised way across a network. DLT is designed to function without a trusted authority. Thus, DLT is a shared database system synced across multiple computers where data updates together without a central boss. The best-known type of distributed ledger is blockchain.
4	Smart contract	Smart contracts are self-executing digital programs stored directly on a blockchain that automatically run when pre-determined conditions are met. They eliminate the need for a central middleman or trusted third party by converting legal or logistical agreements into immutable, mathematically verifiable lines of code.
5	Stablecoins	Stablecoins are a specific class of cryptocurrencies designed to maintain a steady exchange value. They bridge the gap between traditional fiat currencies and DLT by pegging their price 1:1 to an external reference asset—most commonly the USD. *Example*: USDT (Tether) and USDC (USD coin)
6	Token	A token is an object that represents another object (either physical or virtual), or an abstract concept. *Example*: Fungible Tokens (interchangeable, like identical currency notes) and Non-Fungible Tokens (NFTs) (completely unique, like a specific house deed).
7	Tokenisation	Tokenisation means recording an existing asset on a blockchain platform and in order to improve processes around trading and transfer of the asset.

Sr.	Term	Description
8	Web2	Web2 is the current version of the internet characterised by centralised, user-generated content and social media platform monopolies
9	Web3	Web3 refers to a decentralised online ecosystem based on the blockchain where platforms and apps are not owned by a central gatekeeper, but rather by users, who will earn their ownership stake by helping to develop and maintain those services.

Sources:

i) BIS: https://www.bis.org/publ/qtrpdf/r_qt1709y.htm
ii) ECB: https://www.ecb.europa.eu/paym/dlt/html/index.en.html
iii) Worldbank: https://openknowledge.worldbank.org/server/api/core/bitstreams/5166f335-35db-57d7-9c7e-110f7d018f79/content
iv) European Parliament briefing document: https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757580/EPRS_BRI(2023)757580_EN.pdf

Annex-2: Taxation of Virtual Digital Asset (VDA) in India

In India the Virtual Digital Assets (VDAs) are legally defined under Section 2(47A) of the Income Tax Act, 1961. They broadly encompass crypto-assets, Non-Fungible Tokens (NFTs), and any other digital assets, excluding Indian currency, CBDCs, Foreign currency, and notified digital assets. As per the Income Tax Act, income from the transfer of VDAs is computed without any deduction except for the cost of acquisition and the income is taxed at a flat rate of 30

Meaning of virtual digital asset

Virtual Digital Asset, as per the Act, includes:

(a) Information or code or number or token generated through cryptographic means [Section 2(47A)(a)];
(b) Notified Non-fungible token (NFT) [Section 2(47A)(b)];
(c) Any other notified digital asset [Section 2(47A)(c)]; and
(d) Any crypto-asset being a digital representation of value that relies on a cryptographically secured distributed ledger or a similar technology to validate and secure transactions [Section 2(47A)(d)].

and it excludes:

(a) Indian Currency;
(b) Central Bank Digital Currency (CBDC);
(c) Foreign Currency;
(d) Physical NFT [Notification No. 75, dated 30-06-2022];
(e) Gift Card or Vouchers [Notification No. 74, dated 30-06-2022];

(f) Rewards Points or Loyalty Cards [Notification No. 74, dated 30-06-2022];
(g) Subscription of Websites or Applications [Notification No. 74, dated 30-06-2022]; and
(h) Notified Digital Assets.

Tax computation on the transfer of VDA [Section 115BBH]

The income arising from the transfer of virtual digital assets is computed as under:

	Particulars	**Amount (Rs.)**
A	Full Value of Consideration	XXXX
B	*Less: Cost of Acquisition (if any)*	*(XXXX)*
C	**Taxable Income (A – B) (Capital Gains/Business Income)**	**XXXX**

The income from the transfer of VDA is taxed at the rate of 30% without deduction of any expenditure (other than the cost of acquisition, if any) or allowance or set-off of any loss.

TDS on the transfer of VDA [Section 115BBH]

A 1 per cent TDS (tax deduction at source) is applied under Section 194S of the Act when payment is made to a resident for the transfer of VDA. If the payment is made to a non-resident, TDS is deducted under Section 195. No TDS is required if the payment does not exceed Rs. 50,000 (if the payer is an individual/HUF with a business turnover below Rs. 1 crore or professional receipts below Rs. 50 lakh or not having any business/profession income) and Rs. 10,000 (for all other payers).

Reporting of crypto-asset transactions [Section 285BAA]

The Act also requires that specified reporting entities must report transactions involving crypto-assets (as defined in Section 2(47A)(d)) in a prescribed form and timeline.

Source: `https://www.incometaxindia.gov.in/w/taxation-of-virtual-digital-asset-vda-`

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 191-200

Financial Statistics in India: Current Contours and Future Challenges

Partha Ray*

Director, National Institute of Bank Management, Pune

Received: 14 August 2026; Revised: 01 September 2026; Accepted: 10 September 2026

Abstract

This paper examines the architecture and challenges of financial statistics in India, arguing that while finance now constitutes a major share of India's gross value added, the statistical measurement of the financial sector remains somewhat fragmented. The paper describes how India's financial data collection is deliberately distributed across sector-specific regulators— the RBI for banking, SEBI for capital markets, IRDAI for insurance, PFRDA for pensions, and others— each operating within distinct taxonomic frameworks and publication schedules. The paper identifies six critical challenges: (a) financial innovation, particularly fintech and cryptocurrency; (b) the rise of non-banking financial companies and informal lending; (c) data gaps on household balance sheets and micro, small, and medium enterprises; (d) implementation of the Digital Personal Data Protection Act of 2023; (e) increasing cross-border financial integration; (f) the emergence of green finance. Finally, artificial intelligence and alternative data sources present both opportunities and risks, including algorithmic bias and opacity. The paper concludes that robust financial statistics require greater inter-institutional coordination, modernized data governance frameworks, and systematic investment in statistical infrastructure to effectively measure an increasingly sophisticated financial system.

*This article is based on a talk delivered at the Session on Financial Statistics in the 28th Annual Conference of the Society of Statistics, Computers and Applications, on Innovative Trends in Statistical Science, Data Analytics and Interdisciplinary Research-2026 (ITSSDAIR-2026), co-hosted by the Department of Statistics and Centre for Advanced Studies, Savitribai Phule Pune University, on January 5, 2026. The author is indebted to Dr. R.B. Barman, Dr. A.C. Kulshreshtha, Dr. Ajit Joshi, and Dr. Praggya Das, for their comments on the talk. The usual disclaimer applies.

Key words: India; Financial Statistics; RBI; SEBI; Artificial Intelligence.

Corresponding Author: Partha Ray
Email: pray@nibmindia.org

1. Introduction

The architecture of financial statistics in any country perhaps reflects both the institutional maturity and complexity of the financial sector. India is no exception. Any serious engagement with financial data in India, thus, must grapple with a fundamental premise: finance is not merely ancillary to the economy, but is core to economic activity. Interestingly, economists have diametrically opposite views on the issue of direction of causality between finance and growth. While British economist Joan Robinson famously said that "finance follows enterprise," Nobel Laurette US economist Merton Miller argued that financial markets actively could spur real economic activity.

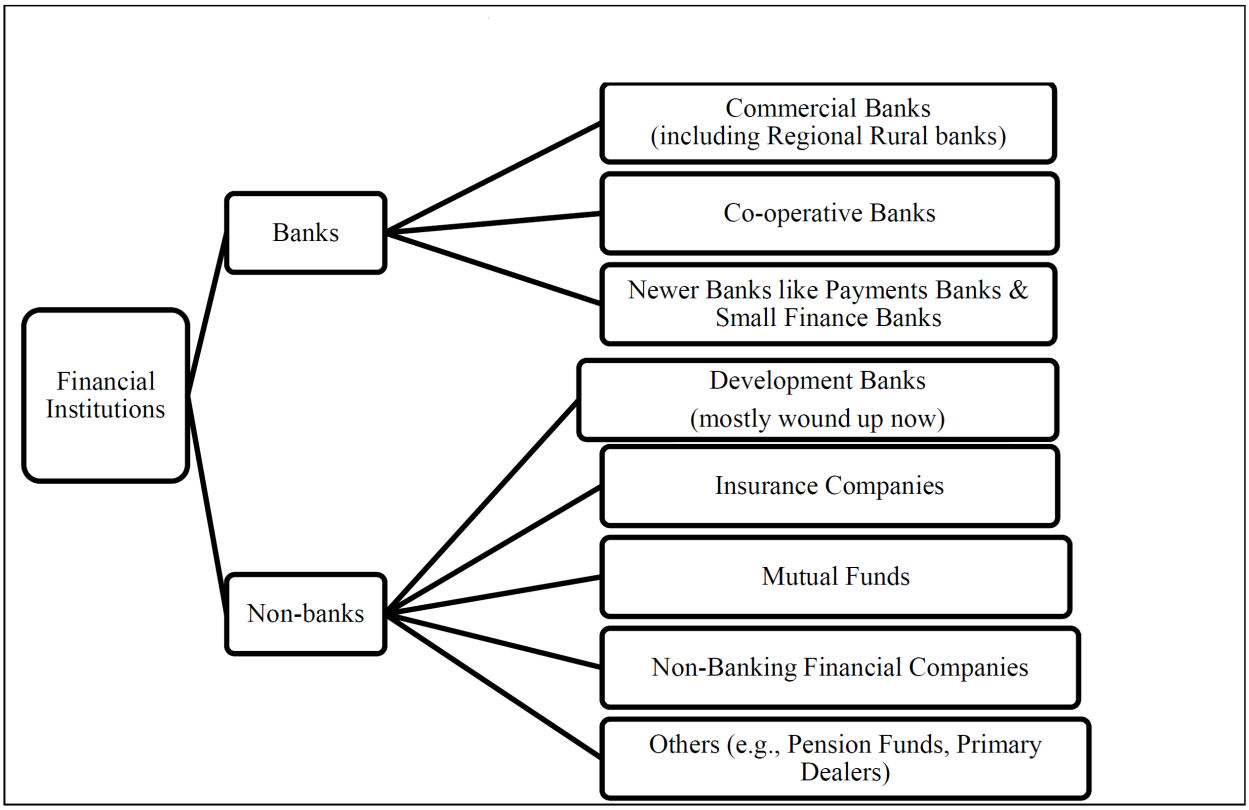
In terms of recent Indian specifics, the "financial, real estate, and professional services" sectors together constitute nearly 24 percent of India's gross value added, a figure that has grown steadily from approximately 19 percent in 2011-12. This expansion underscores the importance of finance in the national economy - both as a component of output and as a critical infrastructure for monetary policy, financial stability, and resource allocation across the economy.

Has this prominence of finance in economic life has not been matched by corresponding clarity, consistency or excitement in the collection, standardization, and dissemination of financial statistics? Do the real sector statistics compilers have skepticism of their financial sector counterparts and vice versa? What are the challenges facing the financial statistician? The present essay delves into some of these questions.

2. The institutional framework

It may be apposite to begin with a description of the Indian financial system. Figure 1 depicts the major constituents of Indian financial system. India's financial statistical system operates through a deliberately distributed architecture, in which responsibility for data collection and compilation rests with sector-specific regulators. The Reserve Bank of India (RBI) oversees banking statistics, extracting granular information from scheduled commercial banks on balance sheet composition, credit deployment, deposits, interest rates, and asset quality. It also compiles data on non-banking financial companies (NBFCs), an increasingly significant component of the financial system. The Securities and Exchange Board of India (SEBI), India's stock market (comprising India's two major stock exchanges, NSE and BSE) regulator, maintains statistics on equity and debt market turnover, market capitalization, IPO and corporate bond issuance, and intermediary activity. The Insurance Regulatory and Development Authority (IRDAI) collects data on life and non-life insurance premiums, penetration metrics, and claims ratios. The Pension Fund Regulatory and Development Authority (PFRDA) oversees statistics on National Pension System subscribers and asset allocation. The Association of Mutual Funds in India (AMFI) publishes assets under management and flow data. The National Payments Corporation of India (NPCI) and RBI jointly maintain statistics on digital payments—UPI, IMPS, NEFT, and RTGS transaction volumes and values. The National Bank for Agriculture and Rural Development (NABARD) compiles agricultural credit statistics.

This distributional model reflects an important principle: the institution most closely engaged with a particular financial domain possesses both the regulatory proximity and technical capability to extract meaningful data from market participants. The RBI's access



Source: Mohan and Ray (2018)

Figure 1: Major constituents of Indian financial system

to supervisory information on bank capital adequacy and liquidity positions, for instance, is not merely statistical convenience but a regulatory imperative. Similarly, SEBI's direct relationship with market intermediaries and exchanges furnishes data that could not be reliably reconstructed from secondary sources[1].

Interestingly, various regulators operate within their own taxonomic frameworks, report at frequencies suited to their particular mandates, and deploy definitions that reflect sectoral custom rather than economy-wide standardization. The RBI publishes its Handbook of Statistics on the Indian Economy and Statistical Tables Relating to Banks in India at frequencies that have evolved with technological change, while the capital markets regulator operates on different publication schedules. Insurance and pension statistics follow their own rhythms. The data format, the compilation procedure, and the presentation often vary.

3. Data collection architecture

The methods through which financial data are extracted and processed reflect the regulatory and market infrastructure of modern finance. Statutory regulatory returns—mandated filings that market participants submit to regulators at specified frequencies—form the foundation. Banks file returns on deposits, advances, investments, and asset quality classifications. NBFCs report on the sectoral composition of their credit deployment. Insurance companies submit premiums and claims data. These returns are not voluntary or survey-

[1]See Schipke *et al.* (2023) for some recent issues on Indian financial sector.

based; they are statutory obligations backed by regulatory authority.

Supervisory reporting systems layer additional detail over these mandated returns. The RBI's enhanced supervisory frameworks require banks to submit granular data on capital adequacy, liquidity coverage ratios, and net stable funding ratios at frequencies that have gradually shortened. Market infrastructure—the stock exchanges themselves, the depositories that hold securities in electronic form, the payment system operators—generate transaction-level data as a byproduct of settlement operations. This data, once the exclusive preserve of regulators, increasingly enters the public statistical domain, though usually in aggregated form.

Survey-based data and administrative records supplement the regulatory returns. The RBI conducts periodic surveys of household and business financial conditions, such as, Consumer Confidence Survey; Inflation Expectations Survey of Households; Survey of Professional Forecasters; Order Books, Inventories and Capacity Utilisation Survey; Industrial Outlook Survey; or Foreign Liabilities and Assets (FLA) Survey – to name a few. The Department of Financial Services tracks credit flows to priority sectors. Regulatory archives contain data on regulatory actions, supervisory corrections, and institutional changes. The challenge in this landscape is not data scarcity but rather the fragmentation of data sources, the heterogeneity of collection frequencies and methodologies, and the administrative burden this heterogeneity imposes on reporting institutions.

Timely data requires reporting at frequencies closer to real time. The days of annual or even quarterly data releases have given way to demands for monthly series, and in some cases, weekly or daily aggregations. The RBI publishes weekly data on monetary aggregates, reserve money, and foreign exchange reserves. The National Payments Council of India (NPCI) releases daily payment system statistics. Data released with a short lag from the reporting period are provisional and subject to revision as fuller information accumulates. A statistic deemed preliminary when released may be substantially revised when more complete information becomes available. The policy-maker or analyst who relies on the preliminary release risks basing decisions on data that diverges significantly from the later-revised figure.

Another challenge is the demands of international standards. Since finance is essentially global, the demands of financial statistics need to consistent with various global standards. Annex 1 gives a description of the data collection architecture in India and various global standards.

4. The challenges being faced by financial statisticians

4.1. Financial innovation

Perhaps no force reshapes financial statistics more than financial innovation. The expansion of fintech, digital lending platforms, and cryptocurrency-based assets has created financial instruments and market mechanisms that sit uneasily within the taxonomic categories devised in earlier eras. When the RBI classifies financial institutions into scheduled commercial banks, cooperative banks, and non-banking financial companies, this taxonomy rests on historical distinctions that turn on licensing, deposit-taking authority, and regulatory jurisdiction. Yet a digital lending platform that extends credit through a technological interface, perhaps without taking deposits in the traditional sense, may not fit neatly into

these categories.

The challenge is not merely one of classification, but of preserving historical comparability while capturing the economic substance of new financial forms. If a peer-to-peer lending platform facilitates credit that would previously have been extended through a traditional bank, the growth in P2P lending may represent a genuine expansion of credit or merely a shift in the channel through which credit flows. Statistical frameworks that treat these channels identically will misrepresent the dynamics of financial activity. Yet frameworks that create new categories for each innovation risk fragmenting the time series and complicating longitudinal analysis.

Illustratively, cryptocurrency and distributed ledger technology present an even more acute variant of this problem. The Central Bank has legitimately articulated reservations about permitting unrestricted crypto-asset trading within India, citing monetary policy concerns and financial stability risks. Yet the emergence of stablecoins, digital rupee initiatives, and cross-border cryptocurrency transactions creates a parallel financial ecosystem that exists outside traditional regulatory perimeters and thus outside conventional financial statistics. The lack of data on cryptocurrency holdings, transaction values, and cross-border flows represents not merely a statistical gap but could turn out to be a potential blind spot in financial stability monitoring.

4.2. Shadow banking, NBFCs and institutional fragmentation

The rise of non-banking financial companies and alternative investment funds represents a conscious policy choice to broaden financial intermediation beyond the commercial banking system. The NBFC sector in India has grown substantially, encompassing housing finance companies, microfinance institutions, infrastructure finance entities, and other specialized lenders. Yet the regulatory boundary between what constitutes a reportable NBFC and what remains outside statistical compilation remains porous. Informal lenders, unregistered microfinance entities, and hybrid organizational forms operate in the interstices of regulation, extending credit without systematic reporting to any central authority.

This institutional fragmentation extends beyond the informal sector into the formal regulatory domain itself. A household's financial relationships may span multiple regulators' domains: deposits with a bank (RBI's purview), equity holdings (SEBI), insurance policies (IRDAI), pension contributions (PFRDA), and credit from an NBFC or financial technology platform (potentially outside systematic reporting). Yet financial statistics rarely aggregate across these domains to present a coherent picture of household financial position. The RBI collects data on bank credit to households; the SEBI publishes data on household equity investments; AMFI reports on mutual fund ownership; but no single statistical source synthesizes this fragmented information into a household balance sheet framework.

4.3. Household and MSME data gaps

The scarcity of high-frequency data on household balance sheets and financial conditions ranks among the persistent weaknesses in India's financial statistics. The RBI conducts the Surveys on Professional Forecasters and Consumer Confidence, providing periodic snapshots of expectations and perceived economic conditions. Yet systematic data on household

financial assets, liabilities, income sources, and wealth distribution remain fragmentary. This gap is particularly consequential because, in India's developing financial system, the household sector serves as a significant credit counterparty—both as a borrower of consumer credit and as a provider of non-institutional credit through informal channels.

The informal financial system remains quantitatively significant in India, particularly in rural areas and among lower-income households. The RBI estimates that informal credit continues to finance a substantial portion of consumption and investment by small and marginal farmers, non-farm entrepreneurs, and other economic agents outside the formal banking system. Yet this credit appears only obliquely in financial statistics, typically as a residual in household saving and borrowing calculations. The absence of systematic data on informal credit flows obscures the true trajectory of financial intermediation and introduces measurement error into assessments of monetary policy transmission.

Micro, small, and medium enterprises (MSMEs) face similar statistical invisibility. While the Reserve Bank collects data on credit flows to the MSME sector from formal lenders, information on balance sheet quality, default rates, and sectoral stress within MSMEs relies heavily on credit registry data from banks. Yet a significant portion of MSME financing occurs through trade credit, retained earnings, and informal borrowing mechanisms. The absence of comprehensive MSME balance sheet data—comparable to what regulatory supervisors maintain for banks and insurance companies—means that assessments of MSME financial health rest on partial information.

4.4. The granularity-privacy tension

As financial statistics have matured and data systems have become more sophisticated, both regulators and researchers have recognized the analytic power of granular, transaction-level data. Rather than analyzing aggregate credit flows to an economic sector, one can examine individual firm credit relationships, allowing for much richer analysis of credit allocation, concentration, and the transmission of monetary policy. Similarly, household-level data permit the analysis of distributional consequences of financial policy in ways that aggregate household statistics cannot.

Yet the movement toward granularity collides with legitimate concerns over privacy and data protection. The Digital Personal Data Protection (DPDP) Act, passed in 2023 with detailed rules notified in late 2025, imposes requirements on data collectors and processors regarding purpose limitation, consent, and security safeguards. The compliance framework specifies that personal data can be collected only for stated purposes and cannot subsequently be used for materially different purposes without fresh consent. The implications for financial statistics are substantial. A dataset initially collected for banking supervision may not be redeployed for monetary policy analysis or research without navigating the DPDP framework.

The DPDP Act establishes a significant compliance date of May 13, 2027, by which regulated entities must redesign their data collection, processing, and storage architectures to align with the act's requirements. The penalties for non-compliance are severe—fines up to ₹250 crore for serious breaches such as inadequate security safeguards or mishandling of children's data. These penalties create powerful incentives for institutions to err on the side of restricting data access rather than risking regulatory action. From a statistical perspective, the DPDP framework raises difficult questions. If anonymization is achieved through

aggregation, then granular data that permits identification of individual entities cannot be shared without consent. Yet consent frameworks that apply retroactively to data collected under earlier regimes create practical complications. A bank that has collected transaction data from millions of customers under the understanding that such data would be available for regulatory purposes may face substantial implementation challenges if it must now obtain fresh consent for each new analytic use.

The challenge becomes particularly acute in the context of cross-border data flows. The DPDP Act permits transfers of personal data to countries whose privacy frameworks the government approves, but the mechanism for whitelisting or blacklisting countries has not been fully clarified. International organizations, standard-setting bodies, and peer central banks who have historically relied on accessing Indian financial statistics may encounter new restrictions.

4.5. Cross-border integration

Financial integration with international capital markets has deepened over the past two decades. Foreign portfolio investors account for a significant share of trading on Indian stock exchanges. Foreign institutional investors hold substantial equity positions in Indian banks and corporations. Indian entities undertake cross-border borrowing in international debt markets. Correspondingly, the measurement of India's financial system and its stability cannot ignore the cross-border dimension.

Yet the standardization of financial data across borders remains incomplete. Different jurisdictions classify financial instruments differently, employ distinct definitional frameworks for capital adequacy, and report at different frequencies. The Basel Committee on Banking Supervision has promulgated common standards that major central banks implement, including India through the RBI. Yet implementation details—the precise definition of eligible capital, the risk weightings applied to different asset classes, the treatments of specific regulatory issues—vary across jurisdictions in ways that can complicate international comparisons.

4.6. Green finance

Additionally, the growing policy emphasis on climate-related risks and sustainable finance has introduced new dimensions to financial data collection for which standardized global methodologies do not yet exist. The Task Force on Climate-related Financial Disclosures (TCFD) has developed a framework for corporate climate risk disclosures, but adoption remains voluntary and variable. The RBI has issued guidance on climate risk management within the banking system and begun collecting data on climate-related lending, but this data collection is still nascent. The classification of "green" financing activities—renewable energy projects, energy efficiency investments, sustainable agriculture—lacks a universally accepted taxonomy despite ongoing international efforts toward standardization.

4.7. Artificial intelligence (AI) and alternative data

The rise of machine learning and large language models has begun to reshape the possibilities and risks in financial statistics. Regulators and academic researchers increasingly

deploy alternative data sources—transaction-level payment information, high-frequency market microstructure data, satellite imagery, social media sentiment—to construct financial indicators that complement or partially substitute for traditional statistical series. These alternative approaches can sometimes yield insights or identify economic phenomena faster than conventional data collection.

The integration of AI-driven statistical methods introduces substantial new challenges. Large language models, by their nature, operate as "black boxes" in which the relationship between inputs and outputs remains opaque even to their developers. A model trained on historical financial data may embed biases present in that data, propagating those biases into the statistics derived from its outputs. The phenomenon of "hallucinations" in LLMs—the generation of plausible-sounding but factually incorrect information—presents particular risks when applied to financial data where accuracy is critical.

The ethical implications warrant serious consideration. The deployment of machine learning algorithms in financial decision-making raises questions about the social acceptability and distributional consequences of automated decision-making. If credit allocation increasingly occurs through algorithmic processes rather than human judgment, and if those algorithms exhibit systematic bias, the result may be discrimination in credit availability that violates both legal norms and principles of financial inclusion. Transparency concerns do remain. The intellectual property claims that attach to proprietary algorithms, the commercial incentives to maintain algorithmic secrecy, and the technical complexity of explaining machine learning decisions all work against the kind of transparent, reproducible statistical practice that has been the aspiration of the statistical profession.

5. In lieu of conclusions: Toward stronger coordination and modern frameworks

Addressing these challenges requires movement on multiple fronts.

First, the various regulatory agencies compiling financial statistics would benefit from more systematic coordination on definitional frameworks and classification standards. While the RBI's role as primary central bank for much of financial statistics provides a natural convening function, the growth of the NBFC sector, the increasing significance of capital markets, and the rise of fintech platforms mean that SEBI, IRDAI, PFRDA, and other regulators are now collecting material financial data. Harmonizing definitions, establishing common data standards, and coordinating on publication frequencies would reduce burden on reporting institutions and improve consistency across statistical domains.

Second, the architecture of financial data governance requires examination in light of the Digital Personal Data Protection Act and the legitimate concerns over privacy and cybersecurity. The RBI and other regulators must design data governance frameworks that balance the need for granular data—essential for effective financial stability monitoring and monetary policy—with the protection of personal information and the security of sensitive financial data. This might involve distinguishing between data required for regulatory purposes and data available for research or broader dissemination, establishing clear protocols for consent and data sharing, and investing in cybersecurity infrastructure proportional to the sensitivity of the data maintained.

Financial statistics matter because financial institutions and markets have become central to resource allocation, risk management, and monetary policy transmission in the Indian economy. Yet the statistical measurement of finance remains a work substantially in progress, populated by sophisticated institutional arrangements alongside notable gaps and inconsistencies.

The challenges identified in this talk —the revisionist timeline of data releases, the arrival of financial innovations that outpace statistical categories, the growth of non-banking finance, data gaps on households and small enterprises, the friction between privacy protection and statistical granularity, the complexities of cross-border financial integration, and the risks embedded in the application of artificial intelligence to financial measurement—are not aberrations in an otherwise smoothly functioning system. Rather, they reflect the inherent difficulty of measuring a complex, dynamic, and increasingly sophisticated financial system while balancing competing imperatives: timely data for policy-making, accuracy and stability in statistical series, protection of sensitive financial information, burden management on reporting institutions, and coordination across multiple regulatory agencies.

For statisticians engaged in financial measurement, the task ahead is both intellectually demanding and institutionally consequential. Robust financial statistics are vital not merely as technical artifacts but as essential infrastructure for financial stability, effective monetary policy, and the efficient allocation of capital in an economy increasingly dependent on sophisticated financial intermediation. The coming years will require greater coordination, ongoing investment in modern data infrastructure, and continued attention to the methodological challenges that arise as finance evolves.

Professor Ronald Coase, Nobel Laureate in Economics in 1991, reportedly said, "If you Torture the data enough, it will confess to anything." The key task of the financial statistician is to get the confessions of the data without any torture. And that is a difficult task.

References

Mohan, R., and Ray, P. (2018). Indian financial sector: Structure, trends and turns. In Ulrich Volz, Peter J. Morgan and Naoyuki Yoshino (ed.s): *Routledge Handbook of Banking and Finance in Asia*, London, Routledge.

Schipke, A., Jarkko, T., Nada, C., and Gulde, A. M. (2023). *India's Financial System: Building the Foundation for Strong and Sustainable Growth.* International Monetary Fund, Washinton DC.

ANNEXURE

Primary official sources on financial statistics

Reserve Bank of India:

- Handbook of Statistics on the Indian Economy (annual and updates)
- Statistical Tables Relating to Banks in India (published series)

- Financial Stability Report (half-yearly)
- RBI Bulletin (monthly; current data and policy analysis)
- RBI's reports on Monetary Policy (quarterly statements and supporting documents)
- Report on Trends and Progress of Banking in India (annual)

SEBI (Securities and Exchange Board of India):

- Annual reports and data on capital markets statistics
- Stock exchange handbooks and periodic market reports from NSE/BSE

IRDAI (Insurance Regulatory and Development Authority):

- Annual reports on insurance sector statistics
- Periodic industry reports on life and non-life insurance

PFRDA (Pension Fund Regulatory and Development Authority):

- Reports on NPS subscriber base and asset allocation
- Periodic disclosures on pension sector statistics

AMFI (Association of Mutual Funds in India):

- Monthly publications on assets under management
- Industry statistics and investor category data

NPCI (National Payments Corporation of India):

- Daily and periodic payment system statistics
- Reports on UPI, IMPS, NEFT, RTGS transaction volumes

NABARD (National Bank for Agriculture and Rural Development):

- Agricultural credit and rural finance statistics

Government of India:

- Economic Survey (annual; Statistical Appendix is particularly relevant)
- Department of Financial Services Annual Reports
- Union Budget documents

Relevant International Standards and Bodies

- Basel Committee on Banking Supervision publications (for capital adequacy and supervisory standards)
- Financial Stability Board reports on shadow banking and non-bank financial intermediation
- Task Force on Climate-related Financial Disclosures (TCFD) framework (for climate-related financial risk measurement)
- IMF and World Bank on financial sector measurement and statistics
- BIS (Bank for International Settlements)

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 201-218

Shrinkage Priors for Bayesian Inference – A Contextual Survey

Anandamayee Majumdar

Department of Mathematics
San Francisco State University, San Francisco, CA-94132, USA

Received: 02 July 2026; Revised: 14 September 2016; Accepted: 15 September 2026

Abstract

Shrinkage priors, alternatively called regularization priors, have become popular in modern Bayesian statistics, particularly in high-dimensional and sparse settings. Their evolution indicates a journey from decision-theoretic insights to structured hierarchical modeling. This survey provides a concise account of foundational contributions, emphasizing the problems each method addressed, the mathematical structure introduced, the practical trade-offs, limitations, and the intuition behind their behavior with practical and recent instances.

Key words: Bayesian model; Hierarchical model; Horseshoe; Lasso; Regularization; Ridge; Shrinkage.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Classical least squares becomes unstable or even inestimable once the number of predictors p approaches or exceeds the sample size n, and it offers no mechanism for exploiting the belief that only a handful of predictors actually matter. Shrinkage priors address both problems at once: by pulling coefficient estimates toward zero, they trade a small amount of bias for a substantial reduction in variance, which very often lowers overall estimation and prediction error relative to unregularized least squares. Bayesian shrinkage priors do not use fixed penalty, rather, a *prior* additionally yields a full posterior distribution rather than a single point estimate, giving coherent, automatic uncertainty quantification without a separate bootstrap or asymptotic approximation. Within the Bayesian framework, continuous shrinkage priors – and the global-local family in particular – are attractive because they combine this variance reduction and built-in inference with a shrinkage profile that adapts per coefficient: most coefficients are pulled sharply toward zero, while a minority of truly large effects can still be estimated with little attenuation, approximating the behavior of discrete variable selection without the combinatorial cost of searching over 2^p models.

Corresponding Author: Anandamayee Majumdar
Email: amajumdar@sfsu.edu

Consider an example, the classical linear model:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{1}$$

When p is large relative to n, the least squares estimator: $\hat{\beta} = (X^T X)^{-1} X^T y$ becomes unstable or undefined. Even when it exists, it can have extremely high variance. Regularization/shrinkage methods address this by introducing bias. Rather than estimating each parameter independently, regularization methods "borrow strength" across parameters.

This theme, first appeared formally in decision theory. Stein (Stein, 1956) showed that for $X \sim \mathcal{N}_p(\theta, I)$, the estimator $\hat{\theta} = X$ is inadmissible for $p \geq 3$. This was indeed shocking, since before this result unbiased estimators were widely considered optimal, no clear reason existed to introduce bias. Stein demonstrated that:

A biased estimator that shrinks toward zero can have strictly lower risk.

We show this with an example: Suppose $p = 10$ and $X_i \sim N(\theta_i, 1)$. The James–Stein estimator $\hat{\theta}_i^{JS} = \left(1 - \frac{8}{\|X\|^2}\right) X_i$ shrinks all coordinates toward zero. The Interpretation of this result implies that (i) if $\|X\|$ is small $\Rightarrow$ strong shrinkage, (ii) if $\|X\|$ is large $\Rightarrow$ weak shrinkage. This adaptive behavior is the conceptual ancestor of modern shrinkage priors.

With the advent of ridge regression, more shrinkage methods were eventually introduced and have been adopted in the Bayesian Statistical Literature. In this overview, the paper is organized as follows: Section 2, 3, 4, 5, discuss the different approaches used for shrinkage in some detail; Ridge regression, the Lasso, Scale Mixtures, and the Horseshoe. Section 6 discusses theoretical properties. Section 7 introduces simulation studies comparing some of these methods in terms of respective MSE and bias performances. We end the paper with a conclusion section.

2. Ridge regression

Ridge regression is a foundational method for stabilizing estimation in the presence of multicollinearity and high-dimensional predictors. Since its introduction by Hoerl and Kennard (Hoerl and Kennard, 1970), it has become a standard tool in statistical modeling, machine learning, and applied sciences.

The ridge estimator is defined as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|_2^2 \right\}. \tag{2}$$

Ridge regression (Hoerl and Kennard, 1970) was introduced to address multicollinearity. This solved the problems of instability when predictors are highly correlated, as well as the exploding variance of least squares estimates.

The Bayesian interpretation of the Ridge regression corresponds to $\beta_j \sim \mathcal{N}(0, \tau^2)$.

Combining likelihood and prior yields a Gaussian posterior distribution:

$$\beta \mid y \sim \mathcal{N}(\mu_{\text{post}}, \Sigma_{\text{post}}), \tag{3}$$

with posterior mean given by:

$$\mu_{\text{post}} = (X^\top X + \lambda I)^{-1} X^\top y, \quad \lambda = \frac{\sigma^2}{\tau^2}. \tag{4}$$

This posterior mean coincides with the classical ridge estimator (Hoerl and Kennard, 1970).

Applications and limitations of ridge regression

This section highlights practical domains where ridge regression has been successfully applied, and where it has limitations.

In genomic prediction problems, the number of genetic markers often far exceeds the number of samples. Ridge regression is particularly well-suited to this setting because it assumes that many predictors contribute small effects rather than enforcing sparsity.

Meuwissen *et al.* (Meuwissen *et al.*, 2001) demonstrated that ridge-type models (often referred to as genomic BLUP) can accurately predict breeding values using dense marker maps. Similarly, de los Campos *et al.* (De Los Campos *et al.*, 2009) showed that ridge regression performs well in genome-wide prediction of quantitative traits. Ridge regression is widely used in genomic selection pipelines because it stabilizes estimation in $p \gg n$ regimes, captures polygenic effects, improves prediction accuracy for complex traits.

Financial datasets often contain highly correlated predictors such as asset returns, macroeconomic indicators, and factor models. Ordinary least squares performs poorly in this setting due to instability in covariance estimation.

Jagannathan and Ma (Jagannathan and Ma, 2003) showed that ridge-type regularization can improve portfolio allocation by reducing estimation error in the covariance matrix. DeMiguel *et al.* (DeMiguel *et al.*, 2009) further demonstrated that regularization/shrinkage-based approaches outperform naive diversification strategies in many empirical settings. Ridge regression is used to stabilize portfolio weights, reduce sensitivity to noisy covariance estimates, improve out-of-sample financial performance.

Ridge regression is a standard baseline in supervised learning due to its simplicity, convexity, and strong generalization properties. It is particularly effective when many features contribute weakly to the response.

Murphy (Murphy, 2012) describes ridge regression as a fundamental regularization method in probabilistic machine learning. It is also closely related to kernel ridge regression, which extends linear ridge regression to nonlinear feature spaces.

Saunders *et al.* (Saunders *et al.*, 1998) showed that ridge regression can be formulated efficiently in dual form, enabling scalable learning in high-dimensional feature spaces. Ridge regression is widely used because it provides strong predictive baselines, handles high-dimensional feature representations, extends naturally to nonlinear models via kernels.

In signal processing and applied mathematics, many estimation problems are ill-posed, meaning that small perturbations in data can lead to large changes in solutions. Ridge regression is equivalent to Tikhonov regularization, which stabilizes such inverse problems.

Tikhonov and Arsenin (Tikhonov and Arsenin, 1977) introduced this framework for solving ill-posed problems, and Hansen (Hansen, 1998) provided a comprehensive treatment of rank-deficient systems and regularization techniques. Ridge regression is used to stabilize ill-posed inverse problems, reconstruct signals from noisy observations, improve numerical conditioning.

Climate datasets often involve highly correlated variables such as temperature, pressure, and atmospheric composition. Ridge regression is useful in this context because it handles multicollinearity effectively. Mann *et al.* (Mann *et al.*, 1998) used regularized regression methods in reconstructing historical climate patterns from proxy data. Schneider (Schneider, 2001) further studied statistical methods for handling incomplete and noisy climate datasets using regularization/shrinkage approaches. Ridge regression helps stabilize long-term climate reconstructions, handle correlated environmental predictors, reduce overfitting in paleoclimate models.

Clinical datasets often contain correlated biomarkers, lab measurements, and demographic variables. Ridge regression is widely used for predictive modeling in medicine because it provides stable estimates and reduces overfitting. Van der Laan *et al.* (Van der Laan *et al.*, 2007) describe ensemble learning frameworks where ridge regression serves as a stable base learner. Steyerberg (Steyerberg, 2019) emphasizes ridge-type regularization in clinical prediction modeling for improving calibration and generalization. Ridge regression is used to: improve risk prediction models, stabilize coefficient estimates in clinical data, support decision-making in healthcare systems.

Text data is high-dimensional and sparse, but ridge regression is often used when dense representations (such as embeddings or tf-idf vectors) are employed. Manning *et al.* (Schütze *et al.*, 2008) describe ridge-type models as effective baselines in information retrieval and text classification tasks. Ridge regression therefore handles large vocabularies efficiently, provides stable classification boundaries, serves as a baseline for more complex NLP models.

Unlike lasso-type methods, ridge regression does not produce sparse solutions but instead distributes shrinkage smoothly across all coefficients, making it particularly suitable for dense, structured signal settings.

Some limitations of ridge regression is that because the ℓ_2 penalty is smooth and strictly convex, the solution shrinks all coefficients toward zero but almost never sets any exactly to zero, so ridge cannot perform automatic feature selection (Hoerl and Kennard, 1970). Also, every coefficient is shrunk by roughly the same proportional factor regardless of its true magnitude, so large signal coefficients are shrunk almost as much as small noise coefficients (Tibshirani, 1996). If the true goal is a parsimonious, interpretable model (*e.g.*, "which 5 of these 200 genes matter"), ridge is the wrong tool — ridge regression is stuck manually thresholding small coefficients after the fact, which has no principled justification and no accompanying uncertainty quantification. In applications where model size matters for deployment cost (*e.g.*, fewer features to collect/measure in a clinical or industrial setting), ridge doesn't help cut anything.

In summary, across applications, ridge regression is most effective when predictors are highly correlated, signals are dense rather than sparse, prediction stability is more important than variable selection.

3. The lasso

Tibshirani (Tibshirani, 1996) introduced the lasso (least absolute shrinkage and selection operator) model:

$$\hat{\beta} = \arg\min \|y - X\beta\|^2 + \lambda \sum |\beta_j|.$$

What was new was that the coefficients can be exactly zero, and the model performs variable selection automatically. A geometric insight is that the ℓ_1 constraint has corners. So the solutions tend to lie on axes, leading to sparsity. From a Bayesian viewpoint the lasso corresponds to:

$$\beta_j \sim \text{Laplace}(0, b).$$

Park and Casella (Park and Casella, 2008) provided a full Bayesian formulation. They showed the Laplace distribution arises as a scale mixture of Normals with an Exponential mixing density. This enables Gibbs sampling, as well as connects shrinkage to hierarchical modeling.

The lasso has been widely adopted across applied fields due to its ability to perform variable selection and regularization simultaneously (Tibshirani, 1996).

Applications and limitations of the lasso

High-throughput genomic experiments produce datasets with thousands of predictors (genes) and relatively few samples. This setting is particularly suited to sparse modeling.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{5}$$

The lasso has been used for gene selection in cancer classification problems, where only a small subset of genes is relevant for distinguishing disease subtypes (Golub *et al.*, 1999; Shevade *et al.*, 2000). This identifies small gene panels from thousands of candidates; improves classification accuracy in high-dimensional settings; enables biological interpretability.

Financial prediction problems often involve a large number of correlated predictors and weak signals. The lasso has been used for forecasting asset returns and selecting relevant predictors in macro-finance models (Fan *et al.*, 2008; Belloni and Chernozhukov, 2011). The lasso can reduces overfitting in noisy data; selects economically meaningful predictors; improves out-of-sample forecasting performance. In Healthcare and Clinical Prediction, electronic health records generate high-dimensional datasets with heterogeneous variables.

The lasso is also used to construct clinical risk prediction models, including disease diagnosis and hospital readmission prediction (Tibshirani, 1996; Wu *et al.*, 2009) where it produces sparse and interpretable models; identifies key clinical risk factors; facilitates deployment in clinical settings.

In neuroscience and brain imaging, functional neuro-imaging datasets contain tens of thousands of predictors (voxels) and relatively few observations. Sparse regression methods,

including the lasso, have been used for decoding cognitive states and identifying relevant brain regions (Yamashita *et al.*, 2008; Ryali *et al.*, 2010). This enables identification of task-relevant brain regions, produces interpretable neural decoding models as well as handles extreme high-dimensionality.

In marketing and customer analytics, modern marketing datasets include detailed behavioral and demographic features. Here, The lasso has been used for customer segmentation, churn prediction, and targeted marketing (Verbeke *et al.*, 2012; Netzer *et al.*, 2012). This identifies key drivers of customer behavior, improves targeting efficiency, reduces model complexity.

In Engineering and Fault Detection, industrial systems generate large volumes of sensor data. Here, the lasso has been used in fault detection and system monitoring by identifying relevant signals among many sensors (Wang *et al.*, 2011; Qin and Goldfarb, 2012). This can detect faults early, improves robustness to noise; enhances system reliability.

Text data leads to extremely high-dimensional feature spaces. The lasso has been used for document classification and sentiment analysis through sparse feature selection (Zhang, 2010; Scott and Varian, 2015) as it Selects informative terms; reduces computational cost; improves generalization performance.

The lasso is also used in signal Processing and compressed sensing. Compressed sensing exploits sparsity to recover signals from limited measurements. The lasso formulation has been widely used for signal recovery in compressed sensing (Candès *et al.*, 2006; Donoho, 2006). This Enables accurate reconstruction from few observations; exploits sparsity in signal representation as well as has applications in imaging and communications.

An up-to-date review of Lasso, and Ridge Regression (Gillariose *et al.*, 2026) covers the adaptive Lasso, and open research directions.

Practical trade-off of lasso

Trade-off 1) Irrepresentable Condition Failure — Sparsity vs. Reliability Under Correlation

The practical advantage of using Lasso, is that when predictors are roughly independent (or the irrepresentable condition genuinely holds), Lasso gives exact support recovery — a clean, interpretable subset of "selected" variables with no need for post hoc thresholding. Also, it is computationally cheap and scalable; a single convex problem with fast solvers (coordinate descent, LARS).

A disadvantage is that in practice — genomics, finance, most observational data — predictors are correlated, so the irrepresentable condition is routinely violated (Zhao and Yu, 2006). Lasso's selected set can then be unstable: small perturbations of the data (bootstrap resamples, different training folds) can flip which variables are selected, even when predictive performance barely changes. Secondly, among a cluster of correlated predictors, Lasso tends to arbitrarily pick one and zero out the rest rather than distributing credit across the group, so "variable X was selected" does not necessarily mean X is causally special — it may simply be the marginal winner within its correlation cluster. This is a genuine problem for scientific inference (*e.g.*, "which gene matters") but is often less consequential for pure prediction,

since the selected proxy variable still captures the correlated signal.

Practical rule of thumb: do not over-interpret *which* variables Lasso selects when predictors are correlated — trust it for prediction, be skeptical of it for causal or scientific claims about individual variables. Use stability selection (Meinshausen and Bühlmann, 2010) or the Elastic Net if more reliable selection under correlation is required.

Trade-off 2) Constant Shrinkage — Simplicity vs. Accuracy on Strong Effects

Another advantage of Lasso, is that a single tuning parameter λ controls both sparsity and shrinkage simultaneously, making cross-validation simple (a one-dimensional search) and the procedure easy to tune and reason about. Further, the Lasso has a predictable, monotonic behavior, increasing λ always shrinks or removes more variables.

A disadvantage of Lasso, is that every retained coefficient is pulled toward zero by the same amount λ, regardless of whether it corresponds to a modest or a large true effect. This systematically underestimates large true coefficients (Fan and Li, 2001), which matters whenever the downstream use case cares about effect size and not merely which variables are selected. This bias does not vanish unless $\lambda \to 0$, but $\lambda \to 0$ also removes the sparsity and regularization benefit — so there is an unavoidable tension: vanilla Lasso cannot simultaneously deliver sparse selection and unbiased large-effect estimates. This is exactly the gap that non-convex penalties (SCAD, MCP) and adaptive/reweighted Lasso variants were designed to close, applying less shrinkage to coefficients already estimated to be large (Zou, 2006).

Practical rule of thumb: If calibrated effect-size estimates are needed (*e.g.*, for reporting inferential confidence, not just prediction), consider a two-stage approach — use Lasso for selection, then refit OLS on the selected subset ("post-Lasso"), or use SCAD/MCP/adaptive Lasso directly (Belloni and Chernozhukov, 2013; Zou, 2006).

Recent work extends the Lasso beyond point estimation and selection toward valid statistical inference in high dimensions.

Debiased (desparsified) lasso

The vanilla Lasso is biased and lacks a tractable asymptotic distribution, so it cannot directly yield confidence intervals. The debiased Lasso corrects this via a bias-correction term based on an approximate inverse of the Gram matrix, giving asymptotically normal estimators for individual coefficients (Zhang and Zhang, 2014; van de Geer *et al.*, 2014). Recent extensions include: 1) Sample-splitting for GLMs: Lasso selects variables on one subsample; debiased Lasso performs inference on a held-out subsample, reducing bias and variance relative to standard MLE (Vazquez *et al.*, 2025). 2) Hidden confounding: the doubly debiased Lasso extends debiasing to settings with unmeasured confounders (Guo *et al.*, 2022). 3) Streaming data: debiasing combined with stochastic gradient descent enables incremental inference without storing the full dataset.

Stability under resampling

Modern theoretical work studies the stability of debiased estimators under data perturbation, linking inferential reliability to resampling-based selection frameworks such as knockoffs (Barber and Candès, 2015) – directly relevant to Lasso's known instability under correlated designs.

4. Scale mixtures

Andrews and Mallows (Andrews and Mallows, 1974) introduced scale mixtures of normals. The idea is to allow the parameter depend on the shrinkage parameter through the variance;

$\beta_j|\lambda_j \sim N(0, \lambda_j^2)$.

Here, λ_j controls shrinkage, and randomizing λ_j creates flexibility. For example, if $\lambda_j^2 \sim$ Inverse-Gamma: $\Rightarrow \beta_j \sim t$.

Heavy tails emerge naturally. For example, Spike-and-Slab (Mitchell & Beauchamp, 1988)'s model given by

$\beta_j \sim (1-\pi)\delta_0 + \pi N(0, \sigma^2)$.

This model solved the problem of explicit variable selection, interpretable inclusion probabilities, though the limitations are that the model was computationally expensive, required combinatorial model space. To address limitations, hierarchical priors were introduced as mentioned in the next section.

Applications of scale mixtures

Scale mixtures of normal distributions form a flexible and widely used class of shrinkage priors in Bayesian statistics. By representing regression coefficients as:

$$\beta_j \mid \lambda_j \sim \mathcal{N}(0, \lambda_j^2), \quad \lambda_j \sim p(\lambda), \tag{6}$$

these models induce heavy-tailed, marginal priors while retaining conditional conjugacy. This structure enables adaptive shrinkage and efficient computation in high-dimensional problems.

Scale-mixture representations underpin many modern shrinkage priors, including Student-t, Bayesian lasso, Horseshoe, and related global-local priors (Andrews and Mallows, 1974; Park and Casella, 2008; Carvalho *et al.*, 2010). This section highlights key applied domains where scale-mixture-based shrinkage priors have been successfully used.

Genomic datasets involve thousands to millions of genetic markers with relatively small sample sizes. Scale-mixture priors are particularly useful because they allow strong shrinkage of noise variables while preserving signals with large effects.

Carvalho *et al.* (Carvalho *et al.*, 2010) demonstrated that Horseshoe-type priors, which are scale mixtures with heavy-tailed local scales, perform well in sparse genetic signal recovery. Bhattacharya *et al.* (Bhattacharya *et al.*, 2015a) further developed Dirichlet-Laplace priors, another scale-mixture construction, for optimal shrinkage in high-dimensional

genomic settings. The practical outcomes are identification of relevant genetic variants in genome-wide association studies; improved prediction of polygenic traits; effective handling of ultra-high-dimensional marker spaces.

Functional neuroimaging data, such as fMRI, involve extremely high-dimensional voxel-level predictors. Scale-mixture shrinkage priors enable sparse and interpretable brain activity models. Yamashita *et al.* (Yamashita *et al.*, 2008) showed that sparse Bayesian regression methods, closely related to scale mixtures, can identify task-relevant brain regions. Ryali *et al.* (Ryali *et al.*, 2010) applied sparse logistic regression models for whole-brain classification, where scale-mixture priors improve robustness and interpretability. Practical outcome include identification of spatially localized brain activity patterns; improved decoding of cognitive states from neural data; reduction of dimensionality in voxel-based regression models.

Financial datasets often contain large sets of correlated predictors with weak signals. Scale-mixture priors provide robustness through heavy-tailed behavior, allowing extreme values to remain non-regularized. Griffin and Brown (Griffin and Brown, 2010a) introduced normal-gamma scale-mixture priors for regression, which have been applied in financial modeling contexts with many predictors. These priors improve stability in portfolio estimation and factor modeling problems along with improved prediction of asset returns in high-dimensional factor models; stabilization of regression coefficients under multicollinearity; better out-of-sample portfolio performance.

Compressed sensing relies on recovering sparse signals from incomplete observations. Scale-mixture priors provide a Bayesian foundation for ℓ_1-type regularization and its generalizations. Candès *et al.* (Candès *et al.*, 2006) and Donoho (Donoho, 2006) established the theoretical basis for sparse signal recovery. Bayesian formulations using scale mixtures, including Student-t and Horseshoe priors, have been widely used for robust reconstruction in noisy environments giving accurate signal reconstruction from under-sampled measurements; robustness to measurement noise. This prior is also useful for its applications in MRI and imaging systems. In modern machine learning, scale-mixture priors are widely used for regularized regression and classification problems. Their ability to interpolate between dense and sparse regimes makes them particularly effective. Polson and Scott (Polson and Scott, 2012) developed a general global-local framework based on scale mixtures, which unifies many shrinkage priors used in predictive modeling. These methods are commonly applied in Bayesian versions of logistic regression and hierarchical models. These models can improve predictive performance in high-dimensional classification tasks; adaptively regularize without manual tuning of sparsity levels; offer efficient posterior computation via conjugate mixtures. Text data produces extremely high-dimensional feature spaces, where scale-mixture priors help balance sparsity and flexibility.Scale-mixture-based Bayesian models have been used for document classification and sentiment analysis, where heavy-tailed priors allow a small subset of informative words to dominate predictions while shrinking irrelevant vocabulary features. Sparse selection of informative words or n-grams provide improved generalization in text classification tasks as well as reduce overfitting in high-dimensional bag-of-words models.

Practical trade-offs of scale-mixture shrinkage priors

1. Computation: scale mixtures avoid the combinatorial search of spike-and-slab priors (Polson and Scott, 2010), but naive Gibbs/HMC samplers scale poorly in high dimensions, requiring specialized fast samplers (Johndrow *et al.*, 2020). 2. Flexibility vs. tuning: the hierarchy $\beta_j \mid \psi_j, \tau \sim N(0, \psi_j\tau)$ unifies the Bayesian Lasso, Horseshoe, and Dirichlet–Laplace priors (Carvalho *et al.*, 2010; Bhattacharya *et al.*, 2015b), but every instance still needs a global scale hyperparameter, and standard defaults are largely conventional rather than principled (Griffin and Brown, 2010b). 3. Structured extensions: the framework extends to dependent data via product representations (Griffin and Hoff, 2024) or spatial sandwich-mixture priors (Nishina *et al.*, 2026), the latter paper builds directly on an earlier, (non-spatial) global-local small-area framework (Tang *et al.*, 2018) but each extension introduces new identifiability and posterior-propriety issues to verify. 4. No exact sparsity: these priors concentrate well near sparse vectors, with Dirichlet–Laplace outperforming the Bayesian Lasso in simulation (Bhattacharya *et al.*, 2015b), but continuous mixing means no posterior draw is ever exactly zero, so "selection" still requires post hoc thresholding (Carvalho *et al.*, 2010).

Scenario	**Consequence**	**Mitigation**
Very high-dimensional p	Slow mixing	Fast samplers (Johndrow *et al.*, 2020)
Choosing global scale	Manual/hyperprior tuning	Well-studied defaults
Correlated/spatial predictors	Dependence ignored	Structured priors (Griffin and Hoff, 2024; Nishina *et al.*, 2026)
Need exact sparse output	No exact zeros	Post hoc thresholding

5. The horseshoe prior

Another noteworthy shrinkage priors is the Horseshoe prior given by the following:

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N}(0, \tau^2\lambda_j^2), \tag{7}$$

$$\lambda_j \sim \text{Cauchy}^+(0, 1), \tag{8}$$

$$\tau \sim \text{Cauchy}^+(0, 1). \tag{9}$$

This prior has two key properties:

1. an infinite spike at zero, encouraging strong shrinkage,
2. heavy tails, allowing large signals to escape shrinkage.

These features make it particularly effective in sparse settings. Later work refined its interpretation and hyperparameter specification (Piironen and Vehtari, 2017a). This is a

continuous analogue of spike-and-slab, namely the Horseshoe Prior (Carvalho *et al.*, 2010) who showed this prior strongly shrinks noise, preserves large signals, behaves like spike-and-slab without discreteness. For example, suppose there are 100 coefficients, 95 are near zero, 5 are large. Then: small coefficients → heavily shrunk, large ones → almost unchanged. This works because the prior has infinite mass near zero, as well as very heavy tails.

Connection of horseshoe priors to spike-and-slab priors

Earlier Bayesian variable selection methods used spike-and-slab priors:

$$\beta_j \sim (1-\pi)\delta_0 + \pi\mathcal{N}(0,\sigma^2), \tag{10}$$

which explicitly model inclusion/exclusion.

Continuous shrinkage priors such as the Horseshoe can be seen as smooth relaxations of spike-and-slab models, avoiding discrete model selection.

Practical trade-off of the horseshoe prior

Strong sparsity vs. unregularized tails: the half-Cauchy local scales give near-spike behavior at zero, but the same heavy tails leave large signals essentially unregularized (Carvalho *et al.*, 2010). This is problematic under weak identifiability (*e.g.* separation in logistic regression), where the likelihood alone cannot constrain large coefficients, causing instability or divergence in MCMC (Piironen and Vehtari, 2017b). The regularized Horseshoe fixes this by capping prior variance via a finite slab scale, trading a small amount of tail flexibility for bounded, stable inference (Piironen and Vehtari, 2017b).

No separate control of sparsity vs. large-coefficient regularization: the standard Horseshoe cannot specify sparsity belief and the regularization of large coefficients independently, which recent robustness studies confirm can still cause numeric instability in MCMC when relying purely on infinite-variance half-Cauchy priors (Fan *et al.*, 2026). The regularized Horseshoe addresses this directly by decoupling the two via an explicit slab width parameter (Piironen and Vehtari, 2017b).

Computational cost at scale: per-iteration MCMC cost for the Horseshoe has historically hindered its use on large-scale data; recent software addresses this with approximate/adaptive algorithms that exploit an effective number of nonzero coefficients to skip negligible ones (Kang and Lee, 2025). The following table gives a summary comparison across models

Method	Sparsity	Bias (large signals)	Adaptivity
Ridge	No	High	Low
Lasso	Yes	Moderate	Medium
Spike-slab	Yes	Low	High
Horseshoe	Yes	Very low	Very high

6. Theoretical properties of shrinkage priors

A central question in high-dimensional Bayesian inference is whether posterior distributions concentrate around the true parameter at an optimal rate.

Posterior contraction

Let β_0 denote the true parameter vector, assumed sparse with s_0 nonzero entries.

Theorem 1 (Posterior Contraction for Sparse Signals)**:** Consider the Gaussian sequence model:

$$Y_j = \beta_{0j} + \varepsilon_j, \quad \varepsilon_j \sim N(0,1).$$

Under a global-local shrinkage prior with sufficiently heavy tails (*e.g.*, Horseshoe), the posterior satisfies:

$$\Pi\left(\|\beta - \beta_0\|_2 \leq M\sqrt{s_0 \log p} \mid Y\right) \to 1,$$

for some constant $M > 0$ as $p \to \infty$.

This can be interpreted as the following: the posterior concentrates at (near) the minimax rate for sparse estimation. The argument of the proof proceeds in three steps:

1. **Prior mass near truth:** Heavy-tailed priors assign sufficient probability to neighborhoods of large signals.
2. **Testing argument:** Construct tests separating β_0 from alternatives at distance ϵ_n.
3. **Control of small coefficients:** The prior places strong mass near zero, shrinking noise coordinates.

Remark 1: The heavy tails of the prior are essential: light-tailed priors (*e.g.*, Gaussian) fail to allocate enough mass to large signals, leading to suboptimal contraction.

More refined results show that the Horseshoe prior adapts to unknown sparsity levels and achieves near-optimal risk without explicit tuning.

Polson and Scott (Polson and Scott, 2010, 2012) developed a general framework for global-local shrinkage priors, which unifies a wide range of Bayesian regularization methods including the Horseshoe, Bayesian lasso, and normal-gamma priors.

Shrinkage factor representation

A key insight of this framework is that posterior shrinkage can be expressed through latent shrinkage factors. Define:

$$\kappa_j = \frac{1}{1 + \tau^2\lambda_j^2},$$

where τ is a global scale parameter and λ_j is a local scale parameter associated with coefficient β_j.

Under Gaussian likelihood models, posterior means admit the approximate form:

$$\mathbb{E}[\beta_j \mid Y_j] \approx (1 - \kappa_j)Y_j.$$

This representation provides a direct interpretation of shrinkage behavior. Figure 3 demonstrates this in the Appendix:

1. $\kappa_j \approx 1$ implies strong shrinkage toward zero (noise-dominated coefficients),

2. $\kappa_j \approx 0$ implies weak shrinkage (signal-dominated coefficients).

7. Simulation comparison: ridge, lasso, and horseshoe

Setup

We compare Ridge (Hoerl and Kennard, 1970), the Lasso (Tibshirani, 1996), and the horseshoe prior (Carvalho *et al.*, 2010) on coefficient estimation error and out-of-sample prediction error under a sparse, correlated design. For each of $B = 30$ replications we generate

$$X_i \sim N(0, \Sigma), \quad \Sigma_{jk} = \rho^{|j-k|}, \quad y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \tag{11}$$

with $n_{\text{train}} = 100$, $n_{\text{test}} = 300$, $p = 100$, Toeplitz correlation $\rho = 0.5$, and a sparse true β with $s_0 = 10$ nonzero entries (magnitude 1.5, random sign and location). The noise variance σ^2 is set so that $\text{Var}(X\beta)/\sigma^2 = 4$ (SNR = 4). All three priors are fit by Gibbs sampling under the same MCMC settings (1500 iterations, 700 burn-in, posterior mean reported), so the comparison isolates the effect of the *prior*, not the fitting procedure.

Models

$$\text{Bayesian Ridge:} \quad \beta_j \mid \tau^2, \sigma^2 \sim N(0, \sigma^2\tau^2), \qquad \tau^2 \sim \text{IG}(\epsilon, \epsilon) \tag{12}$$

$$\text{Bayesian Lasso:} \quad \beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2), \qquad \tau_j^2 \sim \text{Exp}(\lambda^2/2), \qquad \lambda^2 \sim \text{Gamma}(1, 1) \tag{13}$$

$$\text{Horseshoe:} \quad \beta_j \mid \lambda_j, \tau, \sigma^2 \sim N(0, \sigma^2\tau^2\lambda_j^2), \qquad \lambda_j, \tau \sim C^+(0, 1) \tag{14}$$

The Bayesian Ridge shares a single global variance τ^2 across all coefficients (Gibbs updates: Normal for β, Inverse-Gamma for σ^2 and τ^2). The Bayesian Lasso follows Park and Casella (2008), with $1/\tau_j^2$ updated via an Inverse-Gaussian full conditional. The horseshoe uses the Makalic and Schmidt (2016) parameter expansion. Data are generated: $n_{\text{train}} = 100$, $p = 100$, Toeplitz correlation $\rho = 0.5$, $s_0 = 10$ true nonzero coefficients, SNR = 4, $B = 30$ replications.

Results

Method	Coef. MSE	Pred. MSE	Bias (signal)	Bias (null)
Ridge (Gibbs)	0.0979 ± 0.0024	12.55 ± 0.28	0.0228 ± 0.0384	-0.0037 ± 0.0044
Lasso (Gibbs)	0.0982 ± 0.0044	12.89 ± 0.39	0.0288 ± 0.0242	-0.0049 ± 0.0040
Horseshoe (Gibbs)	0.0243 ± 0.0026	7.70 ± 0.31	0.0397 ± 0.0202	-0.0032 ± 0.0020

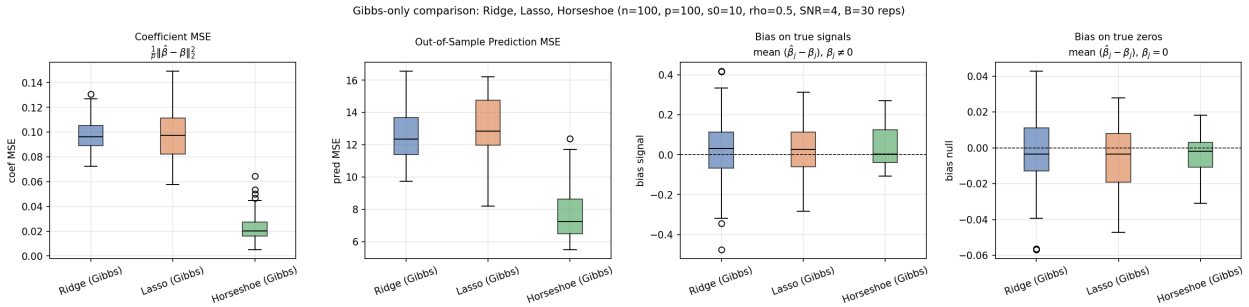


Figure 1: Coefficient MSE, prediction MSE, and bias (on true signal and true null coefficients) across $B = 30$ replications, for the Gibbs-sampled Ridge, Lasso, and horseshoe posteriors.

Discussion: bias vs. variance

Bias on the true null coefficients is negligible and statistically indistinguishable across all three priors (≈ -0.003 to -0.005, well within one standard error of one another): every method correctly shrinks noise coefficients toward zero.

The comparison on true signal coefficients is more subtle. The horseshoe does *not* have the smallest point estimate of bias – at 0.0397 it is in fact the largest of the three – yet it achieves the lowest coefficient MSE. This is reconciled by the classical decomposition

$$\text{MSE} = \text{Bias}^2 + \text{Variance}, \tag{15}$$

together with the horseshoe's markedly smaller standard error on the bias estimate itself (0.0202, roughly half that of Ridge's 0.0384): its posterior-mean estimates are more stable across replications, even though their average deviation from the true signal is not smaller. The horseshoe's MSE advantage over Ridge and Lasso (Gibbs) in this experiment is therefore driven primarily by *variance reduction*, not by reduced bias.

Ridge and Lasso (Gibbs) show similar bias magnitudes to each other (0.023 vs. 0.029) and similarly wide standard errors, consistent with the earlier finding that the weakly-tuned Bayesian Lasso hyperprior on λ^2 causes it to behave more like Ridge than like a sparsity-inducing estimator in this setting.

Given the overlapping standard errors on all three bias estimates, none of the pairwise bias differences would be considered statistically distinguishable at $B = 30$ replications; a larger number of replications or a paired test across replications would be needed before concluding that any one prior has systematically lower bias than another.

8. Conclusion

In summary, we may view shrinkage as the "bias" introduced to reduce variance. The theory introduces heavy tailed distributions as a means of protection for large signals. Some methods introduce hierarchy to include adaptive behavior as in Bayesian hierarchical models. Sometimes continuous priors are introduced to approximate discrete selection.

The evolution of shrinkage priors reflects a progression. First, Stein proposed Ridge regression. Then came the Lasso, followed by Bayesian Hierarchies and finally Global-Local

Priors. Each step addressed a limitation of the previous approach: ridge regression addressed stability, lasso: sparsity, Bayesian methods: uncertainty, horseshoe: adaptivity. Bayesian shrinkage priors now form a central tool in high-dimensional inference, balancing flexibility, interpretability, and theoretical guarantees. Shrinkage priors, also, popularly known as regularization priors, unify ideas from decision theory, optimization, and Bayesian inference. Their ability to adaptively regularize high-dimensional models makes them indispensable in modern statistics as well as complex settings.

References

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B*, **36**, 99–102.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, **43**, 2055–2085.

Belloni, A. and Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation: Stats in the Château Summer School, August 31-September 4, 2009*, pages 121–156. Springer.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015a). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015b). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, **52**, 489–509.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **182**, 375–385.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, **22**, 1915–1953.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J., Zhang, J., and Yu, K. (2008). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *arXiv preprint arXiv:0812.2604*, .

Fan, K., Subedi, S., Dissanayake Pathiranage, V. R., and Wu, C. (2026). Robust Bayesian high-dimensional variable selection and inference with the horseshoe family of priors. *Computational Statistics & Data Analysis*, **219**, 108358.

Gillariose, J., Joseph, J., and Chesneau, C. (2026). Lasso and ridge regression: A comprehensive review of applications and developments in machine learning. *International Journal of Data Science and Analytics*, **21**, 7.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Griffin, J. E. and Brown, P. J. (2010a). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.

Griffin, J. E. and Brown, P. J. (2010b). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.

Griffin, M. and Hoff, P. D. (2024). Structured shrinkage priors. *Journal of Computational and Graphical Statistics*, **33**.

Guo, Z., Čevid, D., and Bühlmann, P. (2022). Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, **50**, 1320–1347.

Hansen, P. C. (1998). *Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*. SIAM.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios. *Journal of Finance*, **4**, 1651–83.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, **21**, 1–61.

Kang, M. and Lee, K. (2025). Mhorseshoe package in R: Approximate algorithm for the horseshoe prior in Bayesian linear model. *SoftwareX*, **31**, 102236.

Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998). Global-scale temperature patterns. *Nature*, **392**, 779–787.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, **31**, 521–543.

Nishina, S., Onizuka, T., and Hashimoto, S. (2026). Global-local shrinkage priors for modeling random effects in multivariate spatial small area estimation. *arXiv preprint arXiv:2601.14752*, .

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Piironen, J. and Vehtari, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Electronic Journal of Statistics*, **11**, 5018–5051.

Piironen, J. and Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**, 5018–5051.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, **9**, 501–538.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.

Qin, Z. T. and Goldfarb, D. (2012). Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, **13**.

Ryali, S. et al. (2010). Sparse logistic regression for fMRI classification. *NeuroImage*, **51**.

Saunders, C. et al. (1998). Ridge regression learning algorithm in dual variables. *ICML*, , 515–521.

Schneider, T. (2001). Analysis of incomplete climate data. *Journal of Climate*, **14**, 853–871.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, volume 39. Cambridge University Press Cambridge.

Scott, S. L. and Varian, H. R. (2015). Bayesian variable selection for nowcasting economic time series. In *Economic Analysis of the Digital Economy*, pages 119–135. University of Chicago Press.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, **11**, 1188–1193.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer, Cham, Switzerland, 2 edition.

Tang, X., Ghosh, M., Ha, N. S., and Sedransk, J. (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, **113**, 1476–1489.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–288.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Winston distributed solely by Halsted Press.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *U.C. Berkeley Division of Biostatistics Working Paper Series*, **222**.

Vazquez, D. L. et al. (2025). Debiased lasso after sample splitting for estimation and inference in high-dimensional generalized linear models. *Canadian Journal of Statistics*, **53**, e11827.

Verbeke, M., Morante, R., Frasconi, P., and De Raedt, L. (2012). A statistical relational learning approach to identify sections in scientific abstracts using sentence and document structure. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*.

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, **5**, 468.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yamashita, O. et al. (2008). Sparse estimation for fMRI decoding. *NeuroImage*, **42**, 1414–1429.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, **76**, 217–242.

Zhang, J. (2010). The influence machine: Nonnegative instance-space learning with differentiated regularization. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 861–866. IEEE.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Special Proceedings: ISBN #: 978-81-950383-3-6
28th Annual Conference, 05-07 January 2026; pp 219-226

Foundations First: How Will Mathematics and Statistics Shape India's AI Ambitions?*

Sudeshna Guha Neogi[1], Abhijit Kulkarni[2], Ramasubramanian Sundararajan[3], Balaji Raman[4] and Sourish Das[5]

[1]*SAS,* [2]*Principal Financial Group,* [3]*SOLUS.ai,*
[4]*Cogitaas,* [5]*Chennai Mathematical Institute*

Received: 03 September 2026; Revised: 05 October 2026; Accepted: 05 October 2026

Abstract

Artificial Intelligence has become the organising ambition of Indian higher education and industry alike. Yet the enthusiasm filling our AI and data science classrooms is thinning the foundational training in probability, statistical inference and mathematics on which those systems rest. This note argues that we mistake fluency with tools for competence in reasoning; that the properties we now demand of AI systems, namely explainability, validation and safety, are statistical properties whatever language we use for them; and that India's national AI ambitions fund the application layer far more generously than the foundations beneath it. We state plainly what we are not arguing, including where machine learning practice has outrun statistics itself. We close with six implementable recommendations, each naming the body that would have to act.

Key words: Statistics education; Artificial intelligence; Curriculum design; NEP 2020; IndiaAI Mission; Statistical consulting.

AMS Subject Classifications: 62K05, 05B05

*This is a short communication and set of recommendations arising from a panel discussion at the Annual Conference of the Society of Statistics, Computer and Applications, Pune, 2026

1. Introduction

When we convened this panel, we set ourselves a deliberately uncomfortable question: if AI is the future, why does the discipline that underwrites it feel like it is losing ground?

The worry is easy to state. Across our universities, students move in large numbers toward applied AI and machine learning courses, while curricula increasingly treat probability theory, statistical inference, linear algebra and optimisation as obstacles to clear rather than foundations to build on. Students respond rationally to what the job market appears to

Corresponding Author: Sourish Das
Email: sourish@cmi.ac.in

reward. In the panel's collective experience they also make a slow-acting mistake, one whose costs will not appear in this year's placement statistics but in the next decade's capacity to do original work.

We should be candid at the outset about the standing of that claim. This note reports the considered experience of practitioners and teachers rather than the results of a systematic study. We return to this limitation, and to what would settle the question, in Section 7.

2. What does AI reliability actually ask for?

The central observation draws a distinction between fluency and understanding. A great many graduates today can do remarkable things: assemble a pipeline, fine-tune a pre-trained model, deploy a working system in an afternoon. These are real skills and it would be foolish to disparage them. But fluency with tools is not the ability to reason about a problem, and the gap between the two is precisely where statistics lives.

As foundations thin, students lose a particular set of intellectual moves: posing a problem as an estimand, so that they know which quantity they are trying to learn before they reach for a method; asking whether the data at hand identify that quantity at all; quantifying uncertainty honestly rather than reporting a point estimate as though it were the truth; distinguishing correlation from causal structure; recognising misspecification and knowing what it does to their conclusions. None of these is exotic. They constitute the ordinary discipline of statistical thinking (Cobb, 2015; Donoho, 2017). But they stay invisible in a workflow that treats held-out accuracy as the only criterion, and they erode when departments demote conceptual courses to optional status. The profession has learned this lesson painfully once already: the American Statistical Association issued a formal statement precisely because practitioners across the sciences had reduced inference to a single number they no longer understood (Wasserstein and Lazar, 2016).

Meng (2018) makes the sharpest available case that this is not a matter of taste. His analysis of the big data paradox shows that a very large but slightly non-representative sample can carry less information than a much smaller random one, and that anyone who reasons about sample size rather than data quality will miss the error entirely. That is a foundational insight, and no amount of tooling substitutes for it.

The strongest objection deserves a hearing. One might argue that a practitioner can always acquire foundations later, when the work demands them, and that meanwhile tool fluency creates genuine value today. There is something to this. But it underestimates how hard it is to retrofit conceptual depth onto someone trained to skip it, and how rarely working life creates the space to do so. Foundations are not a module one bolts on. They form a habit of mind, most cheaply acquired early. A practitioner who has never internalised the question "what am I actually estimating, and with what uncertainty?" is unlikely to start asking it under deadline pressure.

3. What are we arguing?

The deeper point is that the qualities we now demand of AI systems, that they be explainable, validated and safe, are statistical qualities whatever language we dress them in. The panel prefers to demonstrate this rather than assert it, so we name the machinery.

Explainability. Practitioners often present explainability as a layer bolted onto a finished model. In substance it is statistical reasoning: sensitivity of outputs to inputs, identification of high-influence observations, the behaviour of one variable when others stay fixed, and whether the model's assumptions hold at all. Rudin (2019) argues further that for high-stakes decisions the post-hoc explanation of an opaque model is often the wrong goal, and that an interpretable model built for the purpose serves better and loses little or no accuracy. Anyone without a modelling vocabulary cannot make that argument, or evaluate it.

Validation. Validation tells the same story about out-of-sample behaviour. Cross-validation, calibration of predicted probabilities against observed frequencies, and evaluation under strictly proper scoring rules (Gneiting and Raftery, 2007) are not bureaucratic formalities. They carry the scientific content of the claim that a system works. Conformal prediction (Angelopoulos and Bates, 2021) now supplies distribution-free coverage guarantees for the predictions of arbitrary black-box models, including large language models, and stands among the clearest recent demonstrations that a classical statistical idea can answer a modern engineering problem.

Safety. Safety asks about tails. As systems shift toward large generative models, the properties that matter most, namely rare but harmful outputs, behaviour under distribution shift, and drift over time, are properties of probability distributions and of risk. Nobody can construct a serious guardrail without thinking distributionally, which is to say statistically.

Reproducibility and auditability belong on the same list. Both demand a documented chain running from data to estimand to method to conclusion, and both break down when nobody in the room can state what the model estimates.

The erosion of foundations is therefore not a parochial concern of statistics departments protecting territory. It bears directly on whether the next generation can make AI trustworthy, which is the only kind of AI worth deploying in healthcare, finance, public policy and the other domains that operate under heavy regulation.

4. What are we not arguing?

The previous section presents an argument by statisticians that statistics should be compulsory, so we should declare the interest and mark its limits. Three clarifications follow.

We are not defending the inherited syllabus. The case for foundations is not a case for preserving a 1970s statistics curriculum in amber. Many Indian programmes still teach compulsory statistical content as a catalogue of named tests and closed-form distributions, and a graduate who has memorised that catalogue is not thereby prepared for anything. The foundations we mean are probability, linear algebra, optimisation, computation, calibration and causal reasoning, which teachers should deliver with data and code in the room (Nolan and Temple Lang, 2010; De Veaux *et al.*, 2017). Where the classical syllabus obstructs that, departments should cut it rather than defend it.

We do not claim that statistics holds all the answers. Modern machine learning has repeatedly outrun statistical theory. Heavily overparameterised models generalise well in regimes where classical bias-variance reasoning predicts they should not, and the

double descent phenomenon (Belkin *et al.*, 2019) now occupies statisticians precisely because practice found it first. Breiman (2001) warned the profession about this asymmetry a quarter of a century ago and events have proved him substantially right. The traffic between the two fields runs in both directions. Students will not believe a curriculum that casts statistics as the adult supervising an unruly engineering discipline, and they should not.

We are not arguing against AI coursework. Nothing in this note asks for fewer AI and machine learning courses. It asks that they rest on strong foundations.

5. The Indian inflection

This argument carries particular resonance here. India is the country of the Indian Statistical Institute and of a statistical tradition, running through P. C. Mahalanobis, C. R. Rao, R. C. Bose and those who followed, that built large-scale statistical practice into the machinery of national planning long before "data science" was a phrase (Mahalanobis, 1965; Ghosh *et al.*, 1999). It would be a strange irony if the nation that gave the world some of its foundational statistical thinking now produced graduates fluent in calling an API but unsure how to define an estimand.

The present moment opens a door as much as it poses a threat. The National Education Policy 2020 (Ministry of Education, 2020), with its emphasis on multidisciplinary and flexible undergraduate programmes, admits two readings: as a licence to let foundational courses wither into electives, or as room to insist that every student of AI also learns to reason about uncertainty. Boards of studies will settle which reading prevails, course by course, over the next few years.

The national AI programme sharpens the point. The Union Cabinet approved the IndiaAI Mission in March 2024 with an outlay of Rs 10,371.92 crore across seven pillars, covering compute capacity, datasets, applications, startups, safe and trusted AI, and skilling (Press Information Bureau, 2024). Its FutureSkills pillar expands AI course offerings at undergraduate, postgraduate and doctoral level, establishes Data and AI labs in Tier 2 and Tier 3 cities, and routes doctoral fellowships through the top 50 NIRF-ranked institutes (IndiaAI, n.d.). This is a serious and welcome commitment. Our observation is narrow and, we think, uncontroversial: the mission funds the AI layer generously and the mathematical foundations beneath it only incidentally, and no pillar takes responsibility for strengthening doctoral training in statistics and the mathematical sciences.

We would put the underlying asymmetry this way. Compute is capital: expensive, constrained, but purchasable, and India is buying it. No budget line buys the capacity to interrogate a method, recognise when it fails and invent a better one. That capacity grows only from sustained engagement with probability, mathematics and inference, over a decade, in departments that have the people to teach it. It marks the difference between a country that uses AI and one that advances it.

6. Recommendations

The panel held that diagnosis without prescription is cheap. What follows stays deliberately small and implementable, and each recommendation names the body that would have to act.

R1. Make foundations prerequisites, not electives. *For university boards of studies, with AICTE, UGC and NBA.* Every AI and data science degree should require a defined floor of foundational coursework, which students complete before their first deployment-heavy elective. We suggest a minimum of twelve credits spanning probability, statistical inference, linear algebra and optimisation. A board of studies can adopt or contest a number. It can do neither with "adequate grounding".

R2. Institutionalise applied statistical reasoning through a consulting course. *For departments.* A final-year statistical consulting or collaborative problem-solving course of three to four credits, which draws on live problems and pairs an academic supervisor with a practitioner. Such a course teaches what lectures cannot: how to turn a messy question into a formal one, how to communicate uncertainty to a non-specialist, how to defend a modelling choice, and how to take responsibility for a recommendation. The panel regarded this as the single highest-return reform available to us.

R3. Reform assessment, not only syllabus. *For examination boards and course instructors.* Prerequisites examined by rote produce gatekeeping without understanding. At least one third of the assessment weight in foundational courses should rest on open-ended items: formulate the estimand for a described situation, critique a flawed analysis, defend a modelling choice orally.

R4. Build a structured two-way relationship with industry. *For departments and industry advisory boards.* Advisory input into curricula, practitioners co-teaching, faculty engaging with real industrial problems, and industry problems feeding the course in R2. One guardrail: academic bodies retain final authority over curriculum content, so that departments achieve relevance without surrendering the conceptual core to short-term market signals.

R5. Fund the people who carry the foundations. *For IndiaAI FutureSkills, DST, UGC and universities.* Doctoral fellowships in statistics and the mathematical sciences, not only in AI; faculty development programmes for teachers of foundational courses; and recognition for faculty who genuinely bridge theory and practice. Teaching capacity, not policy intent, currently binds R1. Mandating prerequisites without it will produce failure rates rather than understanding.

R6. Carry this discussion into a standing, multi-institution process. *For professional societies, academic associations and curriculum bodies in the statistical, mathematical and computational sciences.* No single panel can settle a question of this scope. These recommendations should be taken up through further panel discussions at conferences and on other academic and industry platforms, and consolidated by a standing curriculum committee drawn from several institutions. Such a process could produce a model syllabus and reading list for the consulting course in R2, faculty development workshops attached to annual conferences, a shared repository of consulting case problems, and a student consulting competition. The audit described in Section 7 would give it an evidence base.

7. Limitations and an invitation

This note reports one panel at one conference. Its central empirical claims, that foundational coursework is losing ground and that the loss carries consequences, reflect the collective professional experience of the panellists, and this note does not establish them with systematic evidence.

They are, however, testable. A curriculum audit of twenty to thirty Indian AI and data science programmes, recording compulsory credits in probability, inference, linear algebra and optimisation together with their position in the degree sequence, would settle the descriptive question within a few months. Programme documents are largely public, the All India Survey on Higher Education (Ministry of Education, 2024) supplies the enrolment denominator, and the AICTE model syllabi (AICTE, n.d.) supply a benchmark. We invite researchers across the statistical and computational sciences to undertake that audit, and we offer this note as the hypothesis it would test. A workable protocol would be the following.

1. **Frame.** Draw a stratified sample of twenty to thirty programmes spanning IITs and NITs, central universities, state universities and private universities, and covering B.Tech in computer science with an AI or ML specialisation, B.Tech in AI and data science, and MSc in data science or statistics.

2. **Record.** For each programme, record total credits, compulsory credits in probability, statistical inference, linear algebra and optimisation, and the semester in which each sits relative to the first deployment-heavy course.

3. **Classify.** Count a course as foundational only when its stated objectives include derivation or proof, so that tool-oriented service courses do not inflate the count.

4. **Benchmark.** Compare against the AICTE model curriculum and against De Veaux *et al.* (2017).

5. **Test displacement, not level.** Where programme documents survive from an earlier year, repeat the exercise at two time points. The hypothesis concerns change over time, and a single snapshot cannot refute it.

6. **Report.** Publish the coding rules and the programme list, report proportions with uncertainty, and describe the sample as a convenience sample unless a complete frame is used.

8. Beyond the hype cycle

AI moves in cycles of enthusiasm; foundations endure across them. Probability, statistics and mathematics have outlasted many waves of fashionable technique and will outlast this one, because they are not techniques but ways of reasoning under uncertainty. The task before this community is to hold institutional nerve: to embrace AI wholeheartedly while refusing to let its excitement hollow out the disciplines that make it reliable.

How well we reason, and not only how much we compute, will decide the future of AI in India. That is a statistical question, and answering it well is our responsibility.

Acknowledgements

We thank the organisers of the SSCA Annual Conference, Pune, 2026 for convening the panel, and the participants from the floor whose questions shaped several of the arguments above. The authors are grateful to the anonymous reviewer whose suggestions have helped in improving the presentation of the manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article. All authors are statisticians and hold professional and institutional interests in the standing of statistics within Indian higher education. We believe the argument stands on its merits, and we record the interest so that readers can judge for themselves.

References

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint* arXiv:2107.07511. https://arxiv.org/abs/2107.07511

All India Council for Technical Education (n.d.). *Model Curriculum for Undergraduate, Postgraduate, Diploma and Minor Segree Programmes in Engineering and Technology*. https://www.aicte-india.org/education/model-syllabus

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, **116**, 15849–15854. https://doi.org/10.1073/pnas.1903070116

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**, 199–231. https://doi.org/10.1214/ss/1009213726

Cobb, G. W. (2015). Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, **69**, 266–282. https://doi.org/10.1080/00031305.2015.1093029

De Veaux, R. D., Agarwal, M., Averett, M. *et al.* (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, **4**, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, **26**, 745–766. https://doi.org/10.1080/10618600.2017.1384734

Ghosh, J. K., Maiti, P., Rao, T. J. and Sinha, B. K. (1999). Evolution of statistics in India. *International Statistical Review*, **67**, 13-34. https://doi.org/10.1111/j.1751-5823.1999.tb00378.x

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378. https://doi.org/10.1198/016214506000001437

IndiaAI (n.d.). *IndiaAI FutureSkills*. Ministry of Electronics and Information Technology, Government of India. https://indiaai.gov.in/hub/indiaai-futureskills

Mahalanobis, P. C. (1965). Statistics as a key technology. *The American Statistician*, **19**, 43-46. https://doi.org/10.1080/00031305.1965.10479715

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, **12**, 685–726. https://doi.org/10.1214/18-AOAS1161SF

Ministry of Education, Government of India (2020). *National Education Policy 2020*. New Delhi: Government of India.

Ministry of Education, Government of India (2024). *All India Survey on Higher Education (AISHE) 2021–22*. Press Information Bureau release, 25 January 2024. https://www.pib.gov.in/PressReleasePage.aspx?PRID=1999713

Nolan, D. and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, **64**, 97–107. https://doi.org/10.1198/tast.2010.09132

Press Information Bureau, Government of India (2024). *Cabinet Approves Ambitious IndiaAI Mission to Strengthen the AI Innovation Ecosystem*. 7 March 2024. https://www.pib.gov.in/PressReleaseIframePage.aspx?PRID=2012355

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, **70**, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Publisher
Society of Statistics, Computer and Applications
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2026

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203
Email: galaxystudio08@gmail.com