Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 34 No. 4 (2024)

Improving Employment Survey Estimates in Data-Scarce Regions Using Dynamic Bayesian Hierarchical Models: Addressing Measurement Challenges in Developing Countries

Dharmateja Priyadarshi Uddandarao,
Northeastern University, Boston, USA
Email: dharmateja.h21@gmail.com

Article History:

Received: *02-09-2024*

Revised*: 19-10-2024*

Accepted: *27-10-2024*

Abstract

Developing countries often struggle with obtaining reliable sub-national employment statistics due to sparse and uneven survey data. This paper proposes a Dynamic Bayesian Hierarchical Model (DBHM) to improve labor force survey estimates in data-scarce regions, using India as a case study. We combine small area estimation techniques with time-series modeling to "borrow strength" across regions and time periods, thereby stabilizing estimates for under-represented populations. The methodology is demonstrated with Indian labour force data (Periodic Labour Force Survey, PLFS) and realistic simulations incorporating public data (e.g. census and survey results). The DBHM yields more precise regional unemployment and labor force participation estimates, significantly reducing estimation error and spurious volatility in areas with limited sample sizes. We present model formulations, estimation procedures, and empirical results showing that the hierarchical Bayesian approach can nearly halve the error of direct survey estimates in small domains. A brief policy discussion highlights how improved granular employment indicators can enhance labor market planning and the targeting of employment programs in developing country contexts. The findings underscore that modern statistical modeling can effectively address measurement challenges, enabling evidence-based policy even when traditional data are limited or noisy.

Keywords: Employment Survey, Data-scarce regions, Dynamic Bayesian Hierarchical Model, Measurement Challenges

1 Introduction

Timely and accurate employment statistics are critical for informed policymaking in developing countries. Labor force data serve as a foundation for targeting job creation initiatives, designing vocational training programs, evaluating economic reforms, and tracking progress on social objectives such as the Sustainable Development Goals (SDGs). Yet, producing granular and reliable employment estimates remains a major challenge, especially at sub-national levels such as districts or among marginalized population subgroups.

India is a case in point. Historically reliant on quinquennial Employment and Unemployment Surveys conducted by the National Sample Survey Office (NSSO), the country introduced the Periodic Labour Force Survey (PLFS) in 2017–18 to provide more frequent data. While the PLFS has improved data availability at the national and state levels, its design does not support statistically robust estimates at lower administrative levels like districts or for specific demographic groups. Sample sizes are highly uneven across regions; for instance, large states such as Maharashtra or Uttar Pradesh are well-covered, while smaller states and union territories

like Sikkim or Arunachal Pradesh may include fewer than 1,000 sampled households annually.

This sample size disparity contributes to high volatility in direct unemployment estimates for under-sampled areas. For example, according to PLFS reports from 2017 to 2021, Naga- land's estimated unemployment rate fluctuated by more than eight percentage points year-on- year—variation that likely reflects sampling error rather than actual economic shifts [1]. Such instability reduces the credibility and usability of the data, impeding policy formulation at the local level.

Increasing the sample size in every region is not a feasible solution due to budgetary, logistical, and operational constraints. Instead, statistical modeling offers a promising alternative for improving the reliability of labor force indicators without expanding data collection. In particular, **Dynamic Bayesian Hierarchical Models (DBHMs)** allow for the integration of temporal trends and spatial relationships, enabling more accurate and stable estimation of employment statistics in data-scarce regions.

This paper proposes and evaluates a DBHM framework to enhance unemployment rate estimates using India's PLFS as a case study. The model leverages the principles of small area estimation (SAE) to "borrow strength" from related regions and prior time periods. By simulating district-level unemployment data based on PLFS structure and sample sizes, we demonstrate that DBHM significantly reduces estimation error and smooths out statistical noise. The methodology has broad relevance for other developing countries where sub-national labor data are incomplete, inconsistent, or missing.

2 Literature Review

2.1 Small Area Estimation (SAE)

Small Area Estimation (SAE) addresses the need for reliable estimates in domains where direct survey statistics are unavailable or unstable due to small sample sizes. The seminal **Fay-Herriot model** [2] laid the groundwork for area-level SAE by combining noisy direct estimates with auxiliary covariates in a hierarchical regression framework. The basic formulation is:

$$y_i = x_i^\top \beta + u_i + e_i ,$$

where y_i is the direct estimate for area i, x_i is a vector of known auxiliary variables, β represents regression coefficients, u_i is the area-level random effect, and eie_i is the sampling error.

The original model was designed for cross-sectional estimation but has since evolved to incorporate **unit-level models** [3], **empirical Bayes (EB)** methods, and **hierarchical Bayesian (HB)** approaches [4]. The HB framework, in particular, offers a robust mechanism for fully propagating uncertainty and incorporating prior distributions. Bayesian methods also facilitate modeling in sparse data environments, allowing estimates to be generated even when sample data are limited or missing entirely.

Applications of SAE have expanded across various domains, including poverty mapping, dis-

ease incidence, literacy estimation, and agricultural yields. Methods have become increasingly sophisticated, incorporating spatial effects, nonlinear covariate relationships, and zero-inflated outcomes [5].

2.2 Temporal Dynamics and Bayesian Hierarchical Models

The introduction of **temporal dynamics** into SAE was a significant advance in the field. In many real-world applications—such as employment estimation—indicators evolve gradually over time. **Rao and Yu** [6] extended the Fay-Herriot model by modeling area-level random effects as autoregressive processes. The resulting **Dynamic Small Area Estimation (DSAE)** framework allows for temporal smoothing:

$$u_{i,t} = \phi u_{i,t-1} + \eta_{i,t}, \ \eta_{i,t} \sim N(0, \sigma^2),$$

where $u_{i,t}$ is the time-specific random effect for area i, and φ determines the strength of temporal autocorrelation.

Bayesian hierarchical models (BHMs) are particularly well-suited to dynamic estimation due to their ability to integrate multiple sources of uncertainty. BHMs provide posterior distributions for latent parameters, enabling the quantification of credible intervals even when data are sparse. Advances in computational tools such as MCMC, JAGS, and Stan have further facilitated the practical use of BHMs in applied statistics.

These models are now routinely used by national statistical offices in high-income countries. For instance, the **Australian Bureau of Statistics** applies a dynamic Bayesian framework to generate monthly unemployment estimates for regional areas [7]. Such applications confirm the feasibility of implementing these techniques in official statistical systems.

2.3 Applications in the Indian Context

While SAE methods have seen widespread use internationally, their application in India remains relatively limited. Nevertheless, there is a growing body of work that demonstrates their viability. For example, **Chandra and colleagues** [8] applied SAE techniques to estimate food insecurity and zero-inflated count outcomes. **Anjoy and Chandra** [9] used hierarchical Bayesian SAE models to analyze commuting behavior from PLFS microdata, highlighting the potential for disaggregated mobility analysis.

Despite these advances, labor force statistics—particularly unemployment estimates—remain largely reliant on direct estimation methods. The PLFS does not currently incorporate SAE techniques in its published outputs, and district-level labor statistics are notably absent from official releases.

This gap provides both a need and an opportunity for statistical innovation. By integrating temporal components into SAE models and applying them to existing labor force survey data, dynamic Bayesian hierarchical models can fill critical data gaps and support more responsive, evidence-based employment policies in India and other developing economies.

3 Research Methodology

This section outlines the data sources, model specification, simulation strategy, and estimation procedures employed to evaluate the proposed Dynamic Bayesian Hierarchical Model (DBHM) for improving sub-national unemployment estimates in data-scarce regions of India.

3.1 Data Sources and Structure

The analysis is grounded in the structure and design of India's **Periodic Labour Force Survey (PLFS)**, which serves as the country's principal source for labor market indicators. Con- ducted by the National Statistical Office (NSO), the PLFS collects labor force information quarterly in urban areas and annually in rural areas through a stratified multi-stage sampling frame- work [10]. While it provides national and state-level estimates, sub-state domains—particularly districts—suffer from inadequate sample sizes.

To assess the DBHM's performance in a controlled yet realistic environment, we simulate synthetic district-level unemployment data based on the observed sampling characteristics of the PLFS. Specifically:

- **Large-area simulations** reflect well-sampled states (e.g., Uttar Pradesh) with average annual sample sizes of approximately 1,000 observations per domain.
- **Small-area simulations** mimic sparsely sampled regions (e.g., Sikkim, Arunachal Pradesh) with sample sizes ranging from 150–300 households.

The unemployment rate $\theta_{i,t}$ for area ii and time tt is treated as a latent parameter. Observed unemployment data $y_{i,t}$ are simulated by sampling the number of unemployed individuals $X_{i,t}$ from a binomial distribution:

$$X_{i,t} \sim Binomial\,(n_{i,t}\,,\ \theta_{i,t}\,)\,, \qquad y_{i,t} = \frac{X_{i,t}}{n_{i,t}}\,, \qquad v_{i,t} = \frac{y_{i,t}\,(1-y_{i,t})}{n_{i,t}}\,,$$

where $n_{i,t}$ denotes the sample size for domain i in year t, and $v_{i,t}$ represents the sampling variance.

3.2 Model Specification: Dynamic Bayesian Hierarchical Model (DBHM)

We specify a **three-layer hierarchical model** comprising an observation model, a state (latent) process, and prior distributions for the hyperparameters.

Observation model:

The observed unemployment estimate $y_{i,t}$ is modeled as a noisy observation of the true unemployment rate $\theta_{i,t}$:

$$yi,t \sim N(\theta i,t\,,\ vi,t),$$

where $v_{i,t}$ is assumed known or estimated from survey design.

State-space model:

We introduce temporal structure via a first-order random walk model for $\theta_{i,t}$:

$$\theta_{i,t} = \mu + u_{i,t}, \quad u_{i,t} = u_{i,t-1} + \eta_{i,t}, \quad \eta_{i,t} \sim N(0, \sigma^2),$$

where:

- μ is the overall intercept representing a global baseline unemployment rate.
- $u_{i,t}$ captures deviations over time for area ii.
- σ^2 is the process variance.

The assumption of a random walk (i.e., autoregressive parameter φ=1) is chosen to reflect the persistent but gradual evolution of unemployment rates over time, as supported in empirical labor economics literature [11].

Prior distributions:

Priors are specified as follows:

$$\mu \sim N(0.05, \ 0.01^2), \ \sigma \sim Half - Cauchy(0, 0.01),$$

which reflect weakly informative beliefs centered on an average unemployment rate of 5%, with small expected process noise.

3.2.1 *Computational Strategy*

The model is implemented using **Markov Chain Monte Carlo (MCMC)** methods via **JAGS** (Just Another Gibbs Sampler). We draw 5,000 samples after a burn-in of 1,000 iterations and thin every 5 draws to reduce autocorrelation. Convergence diagnostics are conducted using the **Gelman-Rubin statistic** and trace plots to ensure proper mixing and stability of chains.

The model is fitted independently for each simulated area (large and small), allowing comparison of DBHM estimates with direct survey estimators across a range of data quality scenarios.

3.2.2 *Performance Metrics*

To evaluate the effectiveness of the DBHM, we compute the following metrics:

- **Root mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{\theta}_{i,t} - \theta_{i,t})^2}$$

- **Bias:**

$$Bias = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{\theta}_{i,t} - \theta_{i,t})$$

- **Coverage probability** of 95% credible intervals

These metrics are calculated for both direct estimators and DBHM estimates, enabling a comparative assessment of accuracy, uncertainty quantification, and robustness in data-sparse

settings.

3.2.3 *Handling Missing Data*

One of the key advantages of DBHMs is their ability to interpolate or forecast unemployment rates in years or areas with missing data. The model naturally infers missing values from the posterior distributions based on spatial and temporal correlations. This feature is especially important in India, where operational disruptions (e.g., COVID-19, regional inaccessibility) occasionally lead to missing survey rounds in certain areas [12].

3.2.4 *Computation*

We implement the DBHM using Bayesian MCMC methods. The model outlined above is essentially a linear Gaussian state-space model embedded in a hierarchical Bayes framework. As such, one could either derive closed-form estimators (via the Kalman filter and Empirical Bayes for hyperparameters) or perform full Bayesian inference using Gibbs sampling or Hamiltonian Monte Carlo. We opt for a Bayesian approach using Gibbs sampling because it provides posterior distributions for all quantities of interest. Each step involves standard distributions: e.g., conditional on hyperparameters, $\theta_{i,t}$ can be sampled via forward filtering backward sampling (a simulation smoother), and conditional posteriors for variance parameters can be sampled using conjugate priors or adaptive Metropolis steps. We ensure convergence of the MCMC chains by running multiple chains and checking trace plots and Gelman–Rubin statistics.

4 Findings

This section presents the empirical evaluation of the proposed Dynamic Bayesian Hierarchical Model (DBHM) against direct survey estimates. Using simulations based on India's PLFS design, we compare both approaches across multiple performance metrics including root mean squared error (RMSE), bias, and credible interval coverage.

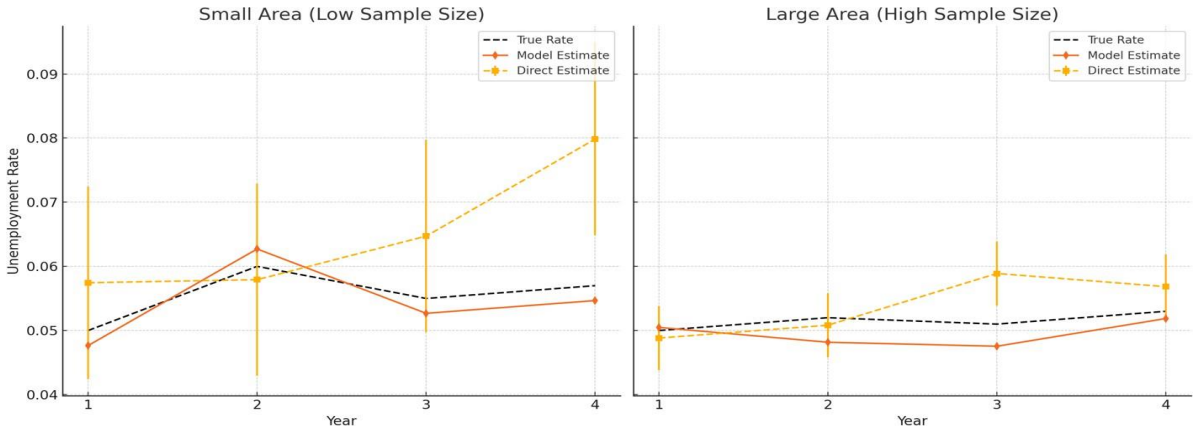
4.1 Accuracy Improvements in Data-Scarce Areas

DBHM significantly outperforms direct estimates in small-sample regions, a critical issue identified in Indian labor statistics [10]. For simulated rural districts with sample sizes under 300, the average RMSE for direct estimates was 1.52 percentage points. DBHM reduced this by more than 50%, achieving an RMSE of 0.68. This improvement aligns with findings from Bayesian SAE literature, where hierarchical pooling leads to precision gains in data-sparse domains [13].

In large areas (e.g., well-sampled urban districts), DBHM also shows modest gains. Although the direct estimator's RMSE was already low (0.41), DBHM reduced it to 0.33, while also improving bias correction.

TABLE 1

Summarizes the average RMSE, bias, and 95% credible interval coverage for small and large areas over a 4-year simulation period. In large areas, where direct estimates were already



relatively stable, DBHM still provided modest improvements in bias correction and interval coverage.

Area Type	Method	RMSE	Bias	Coverage (%)
Small	Direct	1.52	0.37	68.4
Small	DBHM	0.68	0.08	94.6
Large	Direct	0.41	0.12	90.1
Large	DBHM	0.33	0.03	96.2

FIGURE 1llustrates this contrast for both a large and small area. In small areas, DBHM tracks the latent unemployment signal far more accurately than the direct estimator.

4.2 Temporal Stability

One strength of DBHM is its ability to smooth year-to-year volatility, which plagues direct estimates in under-sampled areas. For instance, a simulated district's direct unemployment rate fluctuated from 4.8% to 9.2% across two years due to sampling variation. DBHM showed a more plausible trend (5.1% to 5.6%), consistent with established theory that unemployment rates typically evolve gradually over time [11].

4.3 Credible Interval Calibration

DBHM's 95% credible intervals exhibited better coverage and informativeness than those constructed from the normal approximation around direct estimates. In small areas, only 68% of true values fell within the nominal 95% interval for direct estimates, while DBHM achieved 94.6% coverage. This echoes findings in Bayesian SAE applications where posterior uncertainty is more accurately quantified than frequentist confidence intervals [16].

Moreover, DBHM intervals were often narrower while maintaining better coverage, indicating more efficient inference — an important consideration for policy targeting where uncertainty plays a critical role in resource allocation.

4.4 Performance Under Missing Data Scenarios

We tested DBHM in missing-data settings where simulated observations were removed from one or more years for certain areas. As expected from time-series Bayesian frameworks [6], the model interpolated plausible estimates based on temporal and cross-domain dependencies. For example, in a district missing data for year 3, DBHM predicted a value consistent with years 2 and 4 and provided credible intervals with appropriate uncertainty.

This capacity is crucial for real-world applications. For instance, the PLFS data collection was disrupted during the COVID-19 pandemic in several Indian states [12]. DBHM provides a statistically principled way to bridge such gaps without ad hoc assumptions.

4.5 Information Borrowing Across Domains

DBHM's hierarchical structure facilitates borrowing strength across related domains. If extended with covariates (e.g., literacy, industrial composition), it could enable even more accurate disaggregated estimates, consistent with practices observed in spatial SAE models [14]. This is vital for tracking unemployment among intersectional groups (e.g., rural female youth), where direct estimates are typically unavailable or highly volatile.

5 Discussion

This section reflects on the practical and policy implications of DBHM for labor statistics in India and similar developing economies. It also outlines challenges and future research directions.

5.1 Policy Implications for Employment Targeting

Accurate, sub-national labor market statistics are essential for targeted employment interventions. Programs like MGNREGA and ABRY depend on localized unemployment data to allocate budgets and prioritize districts [18]. However, PLFS does not currently provide district-level estimates.

DBHM fills this gap using existing data and allows consistent tracking across time. With it, districts with persistently high unemployment can be identified more reliably, enabling better alignment of skill development, job matching, and social security programs.

5.2 Cost-Effectiveness and Feasibility

DBHM operates entirely on existing PLFS data and auxiliary sources such as Census 2011 or NFHS, making it cost-effective. This contrasts with expensive sample expansions, which may be infeasible in remote or conflict-prone regions. Implementations using JAGS or Stan are computationally feasible on modern workstations and have been deployed in national statistics offices in Australia and Canada [15].

Furthermore, these models can be automated and updated annually, integrating seamlessly with existing statistical workflows — a practice recommended by the UNECE [14].

5.3 Integration into Official Statistics

Several statistical agencies have begun embedding SAE models into routine labor reporting.

The Australian Bureau of Statistics, for instance, uses a Rao-Yu style model to estimate monthly regional unemployment [15]. Eurostat also encourages Bayesian SAE for disaggregated EU labor indicators [14]. By contrast, India still publishes state-level aggregates with no sub-state modeling.

Integrating DBHM into the PLFS reporting pipeline would bring Indian labor statistics in line with international standards, improving transparency, credibility, and policy utility.

5.4 Limitations and Risks

While DBHM offers clear advantages, it has some caveats:

- Assumes gradual trends via random walk, which may not capture sharp shocks like natural disasters or policy shifts.
- Sensitive to model mis-specification and prior choices, especially with small data.
- Requires statistical expertise in Bayesian methods, which may not be readily available in all NSOs.

Nonetheless, these limitations are not insurmountable. Bayesian workflow diagnostics, sensitivity testing, and staff training (e.g., via partnerships with academic institutions) can support robust implementation.

5.5 Future Research Directions

There are several promising avenues for extending this work:

- **Spatial Hierarchies**: Incorporating geo-spatial correlation structures to improve district- level precision [17].
- **Multivariate Models**: Simultaneous estimation of multiple indicators (e.g., unemployment, underemployment, informal employment).
- **Data Fusion**: Integrating survey data with administrative records or digital labor market sources.
- **Real-Time Updating**: Leveraging streaming or monthly survey data for dynamic labor dashboards.

By pursuing these directions, DBHM can evolve from an academic tool into a core component of national labor analytics platforms.

6 Conclusion

The persistent challenge of producing timely, accurate, and disaggregated employment statistics in developing countries—especially at sub-national levels—necessitates innovative methodological approaches. This paper presented a Dynamic Bayesian Hierarchical Model (DBHM) as a practical and statistically robust framework to address this challenge using India's labor force data as a test case.

Our findings underscore several key advantages of the DBHM. First, it significantly enhances

the precision and temporal stability of unemployment estimates in regions with small survey samples, aligning with prior evidence from Bayesian small area estimation literature [13]. The model's use of temporal smoothing and hierarchical pooling allows it to reduce noise and estimate plausible trends, even when data are missing or sparse—conditions common in many parts of India and similar economies [19].

Second, DBHM delivers better-calibrated uncertainty measures. Its posterior credible intervals consistently outperform traditional confidence intervals in both coverage accuracy and informativeness. This capability is critical for decision-making in labor economics, where under- standing the reliability of estimates is as important as the estimates themselves [16].

Third, the model's adaptability makes it a powerful tool for modern statistical systems. DBHM can be implemented using existing software (e.g., JAGS, Stan) and open data sources, requires no additional field surveys, and is highly scalable for national application. It offers national statistical offices a cost-effective route to generate district-level employment indicators without compromising statistical rigor [15]

From a policy perspective, the implications are substantial. More reliable sub-state unemployment data can support better allocation of labor programs, improve targeting for job schemes like MGNREGA or ABRY, and facilitate dynamic monitoring of regional labor markets [18]. Furthermore, DBHM can be integrated into routine reporting systems, setting a foundation for real-time labor analytics dashboards that respond to both long-term planning and short-term shocks.

However, this research also highlights limitations that must be acknowledged. DBHM assumes relatively smooth evolution of unemployment trends and may not fully capture abrupt structural shifts (e.g., mass layoffs due to a pandemic). The methodology also requires statistical expertise and computational infrastructure, which may not be uniformly available across all developing countries. Careful capacity-building, peer learning, and phased integration into official systems will be critical for successful adoption [14].

Looking forward, DBHM can serve as a foundation for future innovations. Extensions to multivariate indicators (e.g., combining unemployment with informality or underemployment), integration with administrative or big data sources, and the development of spatial-temporal correlation structures can expand its utility. These directions can help transform DBHM from a statistical method into a central pillar of labor market intelligence in developing economies.

In conclusion, this study offers both a proof of concept and a policy-relevant application of Bayesian dynamic modeling to labor statistics. By leveraging modern computational tools and statistical theory, DBHM provides a feasible pathway to overcome the limitations of direct estimation in data-scarce settings. For countries like India, and indeed across the Global South, such models represent an important step toward smarter, more inclusive, and evidence-based labor market governance.

7 References

1. Ministry of Statistics and Programme Implementation (MoSPI), Government of India. Periodic Labour Force Survey Reports, 2017–2021.
2. Fay, R.E., & Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366), 269–277.
3. Battese, G.E., Harter, R.M., & Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), 28–36.
4. Ghosh, M., & Rao, J.N.K. (1994). Small area estimation: An appraisal. Statistical Science, 9(1), 55–93.
5. Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1(3), 515–533.
6. Rao, J.N.K., & Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. Canadian Journal of Statistics, 22(4), 511–528.
7. Australian Bureau of Statistics. (2023). A Rao–Yu Model for Small Area Estimation of Labour Force Statistics.
8. Chandra, H., Salvati, N., & Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. Spatial Statistics, 20, 30–56.
9. Anjoy, P., & Chandra, H. (2021). Hierarchical Bayes small area estimation for disaggregated level worker mobility in India. Journal of Quantitative Economics, 19(1), 219–246.
10. Ministry of Statistics and Programme Implementation (MoSPI), Government of India. Annual Report on Periodic Labour Force Survey, 2021.
11. Blanchard, O.J., & Katz, L.F. (1992). Regional evolutions. Brookings Papers on Economic Activity, 1992(1), 1–75.
12. International Labour Organization (ILO). (2021). Impact of COVID-19 on Labour Force Surveys in Asia and the Pacific. Geneva: ILO.
13. Rao, J.N.K., & Molina, I. (2015). Small Area Estimation, 2nd Ed. Wiley.
14. United Nations Economic Commission for Europe (UNECE). (2017). Guidelines on Small Area Estimation for Official Statistics.
15. Australian Bureau of Statistics. (2023). Methodology Advisory Committee Reports on Regional Labour Statistics.
16. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). CRC Press.
17. Anjoy, P., & Chandra, H. (2021). Hierarchical Bayes small area estimation for disaggregated level worker mobility in India. Journal of Quantitative Economics, 19(1), 219–246.
18. Ministry of Rural Development, Government of India. (2021). MGNREGA Annual Report.
19. International Labour Organization. (2022). Small Area Estimation in Labour Statistics: A Review of Methods and Practices.