Project Report

Aditya Sarkar, Mohammed Zahir Ali

Ashoka University

POL-2094/ECO-3406: Data Science for Social Science Research

Anustubh Agnihotri

May 8, 2023

Introduction

India is a country with a vast and diverse population, with almost 70% of its population residing in rural areas (World Bank, 2023). India's rural economy represents over 50% of its national income, and rural growth and development is a key driver supporting India's overall growth and
Development (Vikash Singh, 2022), thus it represents a very critical area for development and growth. Despite this, inadequate infrastructure, limited access to basic amenities, and low levels of economic development are pervasive problems in rural areas. Consequently, it is important to explore the relationship between infrastructural development and the economic growth factors in rural settlements in India.

The link between infrastructural development and economic parameters areas is very well established in research and papers and its importance cannot be overstated in a developing economy like India. M.S. Bhatia, in his paper titled "*Rural Infrastructure and Growth*" explains how adequate infrastructural support is a prerequisite for accelerated economic development. He claims that many people living in rural settlements do not even have access to the bare minimal infrastructural amenities and for the nation to attain economic stability and growth overall, there has to be a big push to infrastructural developments in rural India (Bhatia, 1999).
Foster and Rosenzweig in their book on "Economic Development and Cultural Change " study whether agricultural development actually leads to economic and income growth within the rural sector of India. They claim that agricultural development is necessary for poverty reduction and development of the non-agricultural sector (Foster & Rosenzweig, 2004). In the paper *"Road Infrastructure Development and Economic Growth"* by Ng et al, it is claimed that Improving basic infrastructural amenities like *pucca* roads, establishment of schools and colleges and technological advancements like telecommunications and electricity will drive higher employment and consumption patterns (Ng et al., 2019). Avinash Kaur and Rajinder Kaur in their paper also make the same claims for the rural settlements in Punjab (Kaur, 2018).

Despite many papers and academic work studying these links, there still exist gaps in terms of size of analysis and depth and range of factors or features considered. Some of the papers

establishing these claims are based on data from the last century and all of them vary in terms of the time of research and thus the datasets used by almost all these papers are different and mostly do not consider data points across the country. Therefore, there is some lack of research and analysis when it comes to analyzing data from a common dataset,and studying the relationship between the multiple parameters of both economic growth and infrastructural development simultaneously in rural settlements across the country. This research paper aims to bridge the gaps by analyzing many factors and proxies for Economic Growth and Infrastructural Development, using the SHRUG dataset which includes rural settlements all across the country as the fundamental granular data points. Research and analysis with more data and depth should produce more reliable observations, facilitating developmental policy making for economic and infrastructural growth of Rural India.

There have been quite a few government policies and schemes focused on boosting the economy and infrastructure of Rural India. Some of the national schemes are Deen Dayal Upadhyay Grameen Kaushal Yojna (DDUGKY) which is a placement linked skill development scheme for rural poor youth. The Mahatma Gandhi National Rural Employment Guarantee Act" (MGNREGA) which is an Indian labour law and social security measure that aims to provide ‘right to work' to the people falling Below Poverty Line. There are many more such schemes even in different states and further understanding of the metrics will help make more such focused policies and schemes targeted towards a particular state or region.

By examining the relationship between these metrics, policymakers can identify the areas where investment is most needed to promote economic growth and development. For example, if the analysis reveals a strong correlation between agricultural employment and infrastructural development, then policymakers can prioritize investments in agriculture-related infrastructure, such as irrigation and storage facilities, to boost economic growth in rural areas. Similarly, if the analysis shows that there is a positive correlation between per capita consumption and the length of roads, policymakers can prioritize investments in road infrastructure to improve access to markets and increase consumption. A more focused planning and policy making for different regions and states can promote sustainable and inclusive development in rural areas by focusing on the most important infrastructural factors first.

Further expanding the scope of the research, this paper aims to analyze how different metrics of economic growth are differently affected by metrics of infrastructural development across different regions. This paper compares these statistics by comparing rural settlements in the under-developed BIMARU states in comparison to the rest of the states in India.

Madhusudan Ghosh in his paper " Infrastructure and Development in Rural India" claims that with poor infrastructure, even a marginal improvement in its quantity and quality could significantly improve economic development and human well-being (Ghosh, 2017).

A paper by Buddhadeb Ghosh and Prabir De explores the role played by infrastructure in determining the level of economic development across the various states over different time spans during the past quarter century.They found that for almost all the metrics considered for economic growth, they were less stable and more significantly affected by the infrastructure development metrics for a select few states consistently. The states being Assam, Bihar, Rajasthan, Odisha in most of the cases. One of the possible reasons stated for this is the fact that these states in the given time period were comparatively under-developed. Our paper, while studying these metrics across BIMARU and other states, expects to find similar results.

Overall, the research paper will thus adopt an exploratory approach, employing quantitative methods to gather and analyze data from various SHRUG datasets. By examining the relationship between employment, consumption patterns, and different proxies of infrastructural development in rural India, the findings of this study will provide valuable insights for policymakers and researchers seeking to promote sustainable and inclusive development in rural areas.

Methodology

The primary data source used in this research is the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). This dataset provides information on socioeconomic development in India at the settlement level. The datasets that we are particularly interested in include the Primary Census (2011), Economic Census (2013), Socio Economic and Caste Census (2012), and Night Lights dataset. These datasets offer valuable insights into population demographics, economic activity, and infrastructure development.

Initially, we intended to include the PMGSY Roads dataset in our analysis, as it provides information on the length of paved roads, which can serve as a proxy for infrastructure development in rural areas. However, since the census data we are working with dates back to 2011 and 2013, we limited our analysis to datasets from the same time period to maintain consistency. Upon closer inspection, we discovered that the PMGSY roads dataset contained a significant amount of missing data points, making it challenging to isolate relevant information. Thus, we decided not to include this dataset in our analysis.

The SHRUG datasets have data points at the level of a SHRID, which represents a settlement that can be a mix of both urban and rural elements but is predominantly rural. To ensure that the SHRIDs we are looking at are indeed rural, we employed a percentage of the rural population threshold. We started with a high threshold of 0.99, which indicated that only 236 SHRIDs did not pass the bar, suggesting that the SHRIDs we are examining are predominantly rural.

For operationalizing our variables, we used the following variables as proxies from the datasets:

- Proxies for Infrastructural Development (to be aggregated)
    - Number of Educational Institutions (Schools and Colleges)
    - Power Supply for domestic, commercial and agricultural usage across the year
    - Night Light Intensity
- Proxies for Economic Development (to be aggregated for analysis)
    - Non-farm Employment

- Agricultural Employment
- Per Capita Consumption
- Per Capita Purchasing Power Parity

To maintain consistency in our analysis, we limited the Economic Census and the Night Lights datasets to only include variables from 2013, which corresponds to the time period of the census data we are analyzing. Additionally, we filtered the datasets to only include SHRIDs that are common across all datasets to ensure that our analysis is based on a consistent sample of rural settlements.

In order to aggregate the metrics for Education, Power Supply, and Night Light, we employed specific methods. For Education, we calculated the total number of educational institutions, including primary schools, secondary schools, middle schools, secondary schools, senior secondary schools, and colleges. We then used their linear sum as the final metric for educational infrastructure development in each SHRID.

For Power Supply, we calculated the mean power supply across summer and winter months for all use cases, including commercial, domestic, and agricultural use. This metric provides insight into the availability and accessibility of electricity in rural areas.

Finally, for Night Light, we divided the total light for 2013 by the number of cells to get an average night light figure for each SHRID. This metric provides a proxy for the level of economic activity and infrastructure development in each rural settlement.

The Economic Census dataset has a limitation in that it only includes data on non-farm employment, which leaves out a significant portion of the population in rural areas that are involved in agriculture. Therefore, in order to account for the importance of agriculture in our analyses, we merged the SECC and PC datasets to create a metric for agricultural employment. This involved calculating the average household size using a new variable called hh_size_mean, which is based on both urban and rural figures. We handled missing values using the ifelse() function, taking the mean when both values were non-missing and using the non-missing value

when only one was available. We dropped rows with missing values for the final hh_size_mean variable to ensure the accuracy of our results.

The agricultural employment variable was then calculated by multiplying the share of households involved in cultivation with the total number of households and the mean household size, and then dividing by two to account for the number of members per household that would be involved in agriculture. Similarly, we derived the per capita consumption and poverty rate (PPP) using the consumption and PPP figures from the SECC dataset by taking the mean of the figures for urban and rural households. Rows with missing values for these new columns were also dropped, and all datasets were subset accordingly. Finally, we added night light figures to this dataset to create the final dataset.

To ensure comparability, we normalized the variables per 1,000 population and generated a correlation matrix. However, despite our efforts, we failed to find any meaningful correlations. This brings into question the validity of our aggregated variables. Although we have referred to other papers that have worked with similar datasets, the formulas used in those studies might not be transferable to our current dataset. Future studies may need to consider alternative methods for operationalizing variables or explore additional data sources to validate our findings.

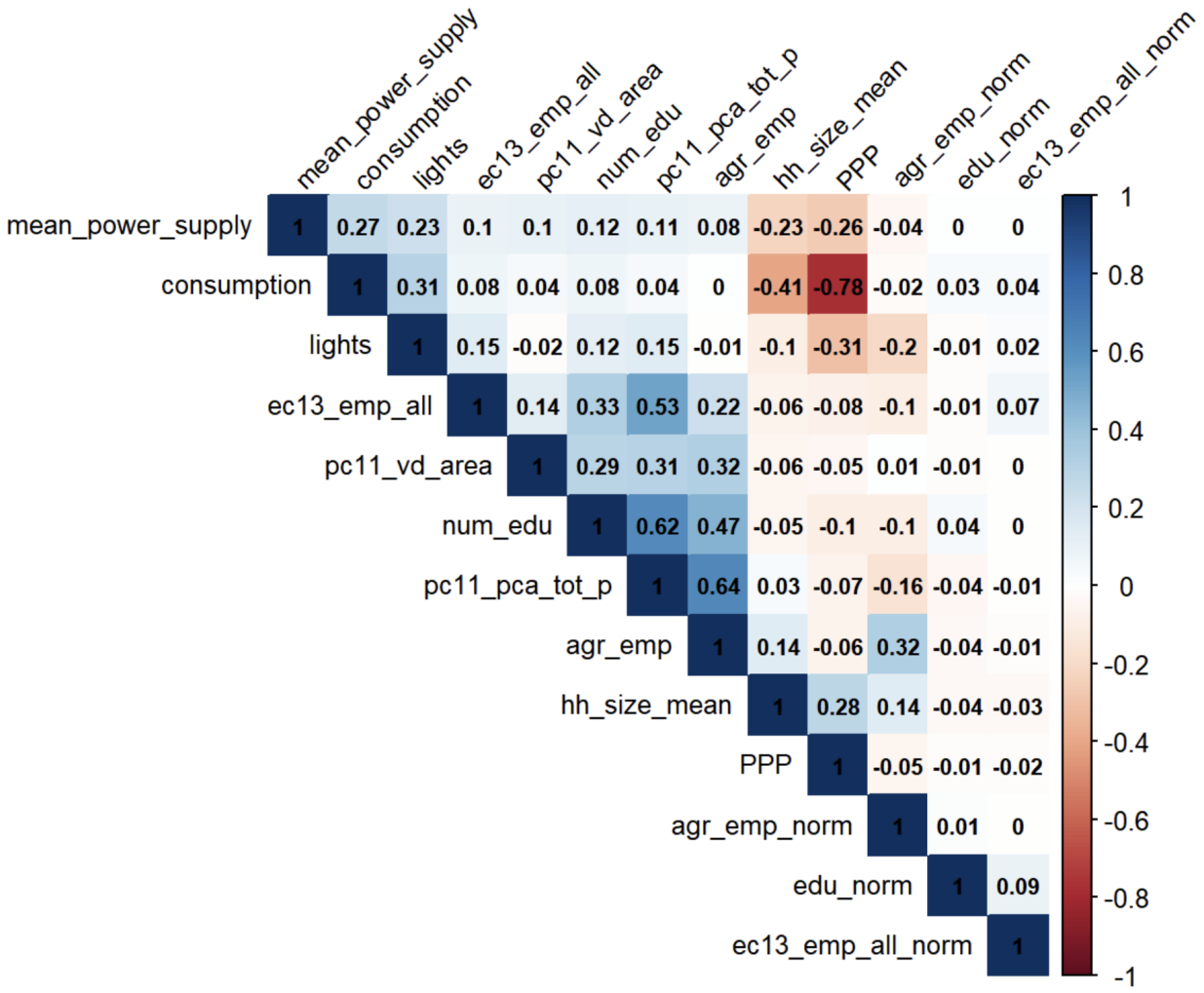

Figure 1. Correlation matrix for aggregated variables

Principal Component Analysis (PCA) is a widely used unsupervised machine learning technique that seeks to reduce the dimensionality of a dataset by identifying the principal components that explain a large portion of the variation in the data. It is an ideal technique for datasets that exhibit high multicollinearity, which can lead to unstable and unreliable estimates in linear regression models. By combining the input variables in a specific way, PCA allows us to drop the least important variables while retaining the most valuable parts of all of the variables. The resulting principal components are orthogonal to one another, which satisfies the assumptions of a linear regression model that require the independent variables to be independent of one another.

In our analysis, we applied PCA to the X variables to aggregate the variables for education and power supply, which had multiple variables for similar metrics and would inevitably lead to multicollinearity. The resulting principal components were a weighted average of all X variables such that the first few PCs explained most of the variance in the dataset. This allowed us to eliminate multicollinearity and reduce the dimensionality of the dataset.

However, while it would have been worthwhile to consider metrics of economic growth as X variables and generate their principal components to see how they are linked to infrastructure, the low number of variables and their low multicollinearity suggest that using PCA for aggregation is not necessary or appropriate. PCA works best for variables that are highly correlated, as it creates orthogonal projections based on the existing relationships within the variables to best explain the variance in the dataset. In this case, the plot generated from the correlation matrix shows that the selected metrics of economic growth are not multicollinear.

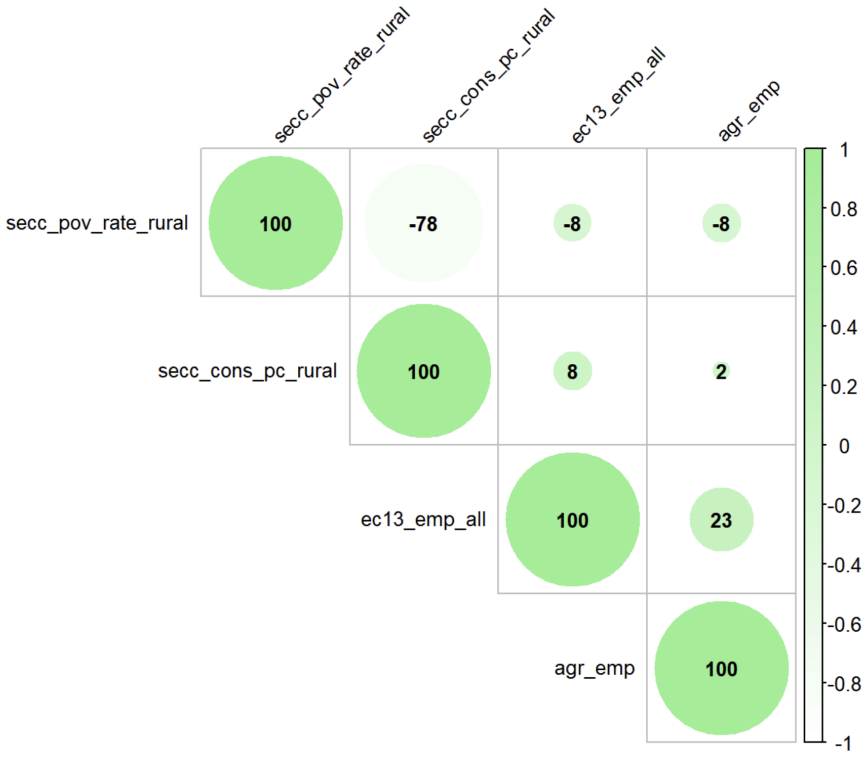

Figure 2. Correlation matrix for metrics of economic growth

One of the most important considerations when using Principal Component Analysis (PCA) is dealing with missing data. In our study, we faced challenges with missing data in the initial dataset. We explored two commonly used methods of imputation: mean-imputation and kNN-imputation. Mean-imputation fills in missing values for a variable with the mean value of all the other data points for that variable, while kNN-imputation groups data points into multiple clusters and then fills in missing values with means of each cluster. However, due to the large size of our dataset (over 500,000 data points), both of these methods proved to be computationally demanding. Moreover, the use of these imputation methods can be controversial, as adding new values may alter the original relationships between the variables. Thus, we made the decision to remove data points with missing values.

After merging the state keys from the SHRUG dataset, we filtered the data to include only the BIMARU states, which consist of Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh.

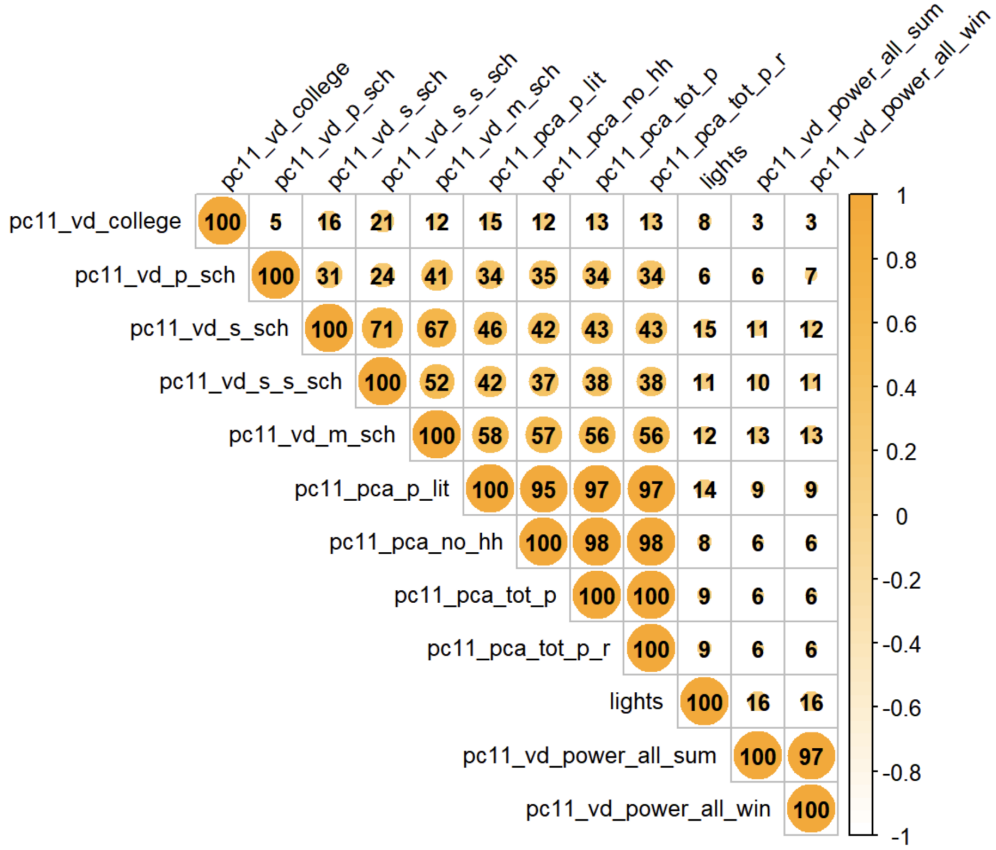

Figure 3.1. Correlation matrix for variables of infrastructure in BIMARU states

We then plotted a correlation matrix to evaluate the dataset for PCA, and found evidence of multicollinearity. In order to proceed with PCA, we scaled the data, fit the factor model, and extracted the eigenvalues, which we used to create a scree plot. This allowed us to identify the Principal Components that explained the most variance in the dataset.

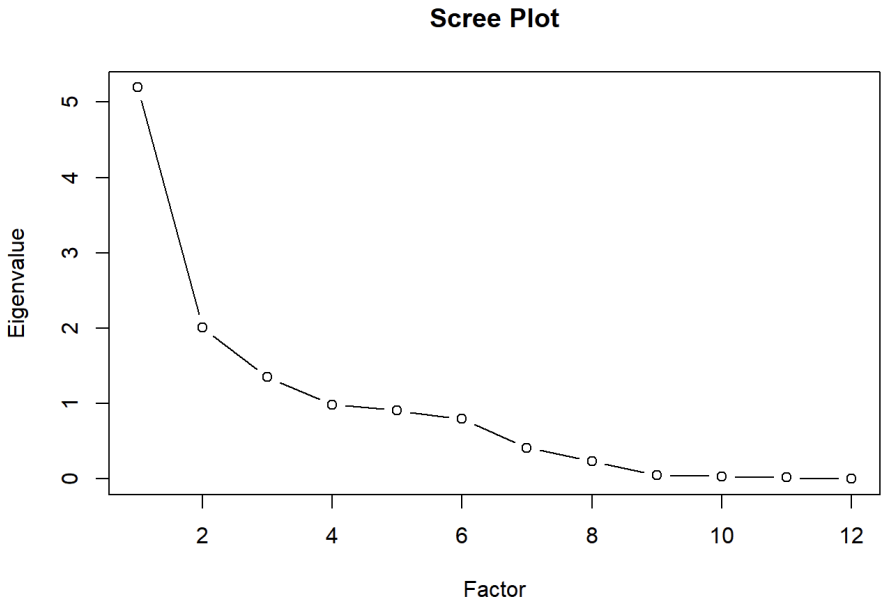

Figure 3.2. Scree Plot for Eigenvectors of BIMARU states variables

PCA generates as many Principal Components as there are variables in the original dataset, but for dimensionality reduction, we need to evaluate which PCs to consider for analysis. This is done by looking only at those PCs which explain most of the variance in the dataset.

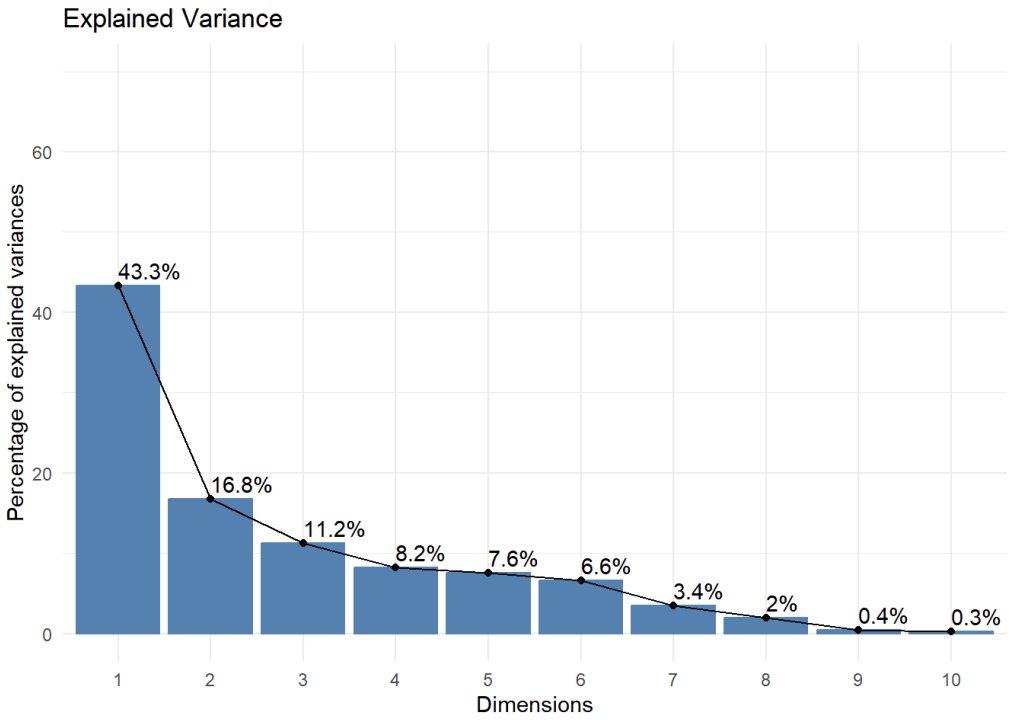

Figure 3.3. Plot for Variance explained by the PCs (BIMARU)

We found that the first four PCs explained almost 80% of the cumulative variance in the dataset, and decided to limit our further analysis to these four PCs. To ensure that these PCs were orthogonal to each other, we plotted another correlation matrix. We then ran multiple linear regressions with scores of these four Principal Components as predictors and metrics of economic growth, including Poverty Rate (Per Capita PPP), Non-farm Employment, Agricultural Employment, and Per Capita Consumption.

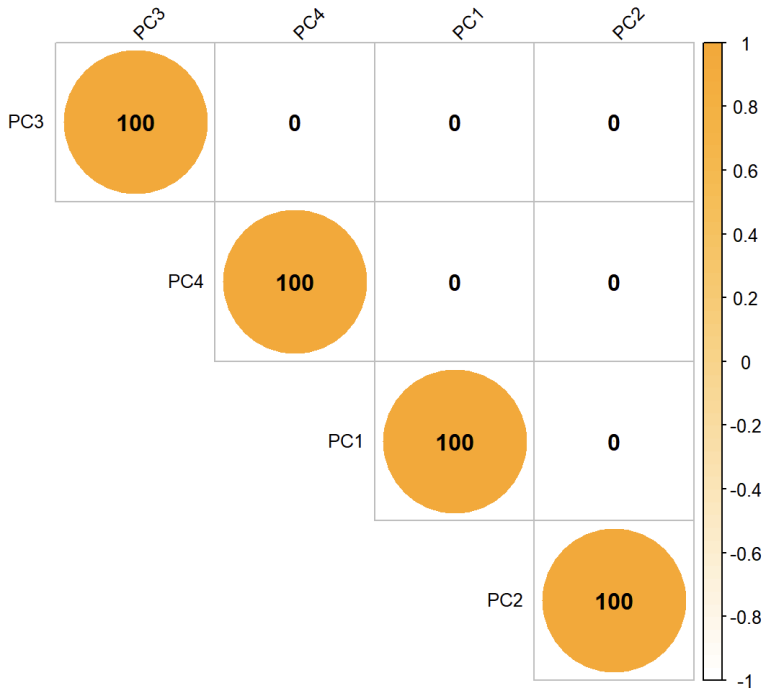

Figure 3.4. Correlation matrix for PCs (BIMARU)

To evaluate the results of our regressions in a broader context, we repeated the process for the entire dataset, filtering out the BIMARU states. We again checked for multicollinearity by plotting a correlation matrix, and found that the first four PCs explained almost 80% of the variance in the dataset. We ran multiple linear regressions with the same predictors and metrics of economic growth to compare the results with those obtained from the BIMARU analysis.

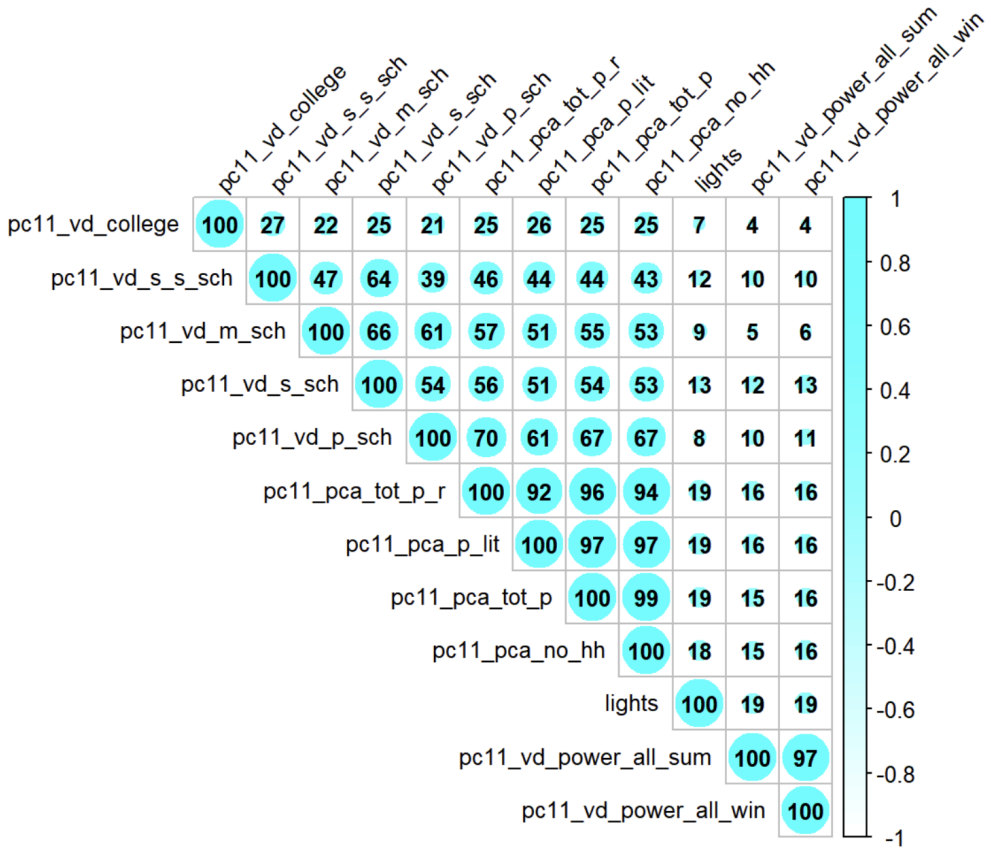

Figure 4.1. Correlation matrix for variables of infrastructure in non-BIMARU states

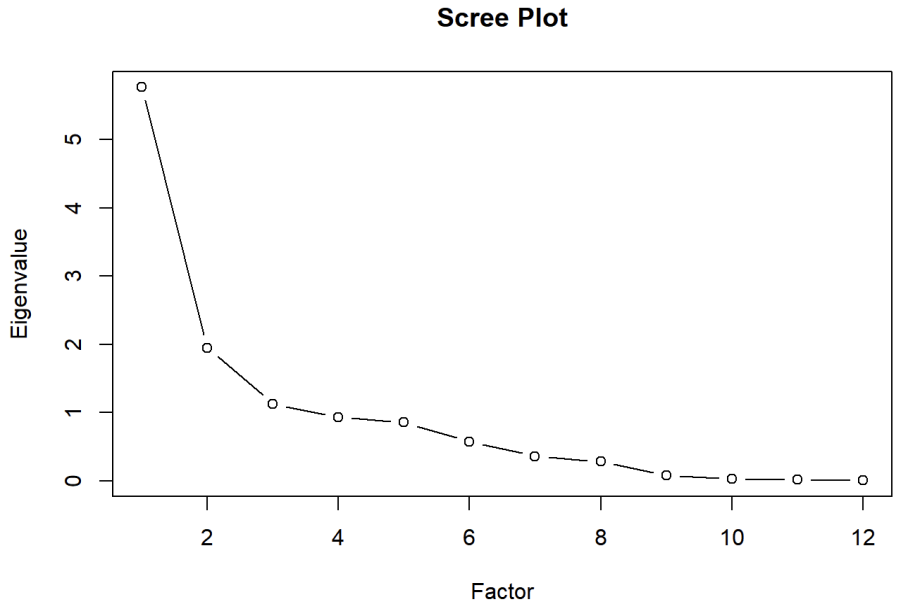

Figure 4.2. Scree Plot for Eigenvectors of non-BIMARU states variables

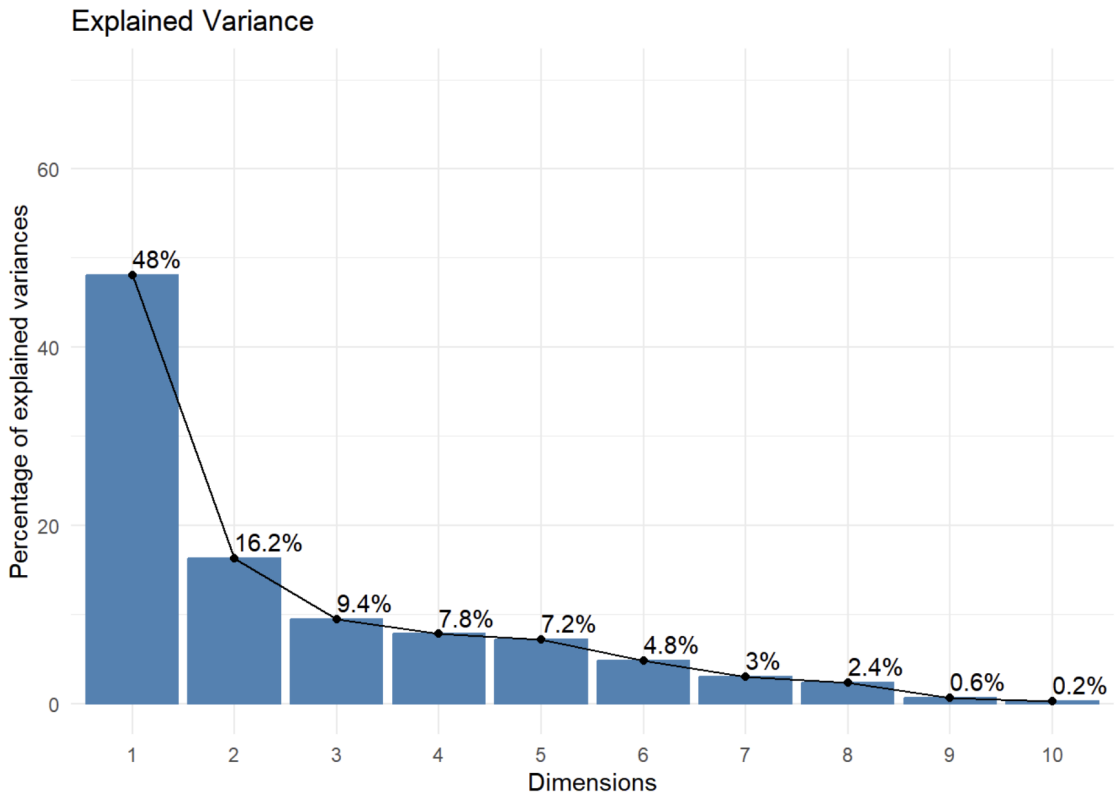

Figure 4.3. Plot for Variance explained by the PCs (non-BIMARU)

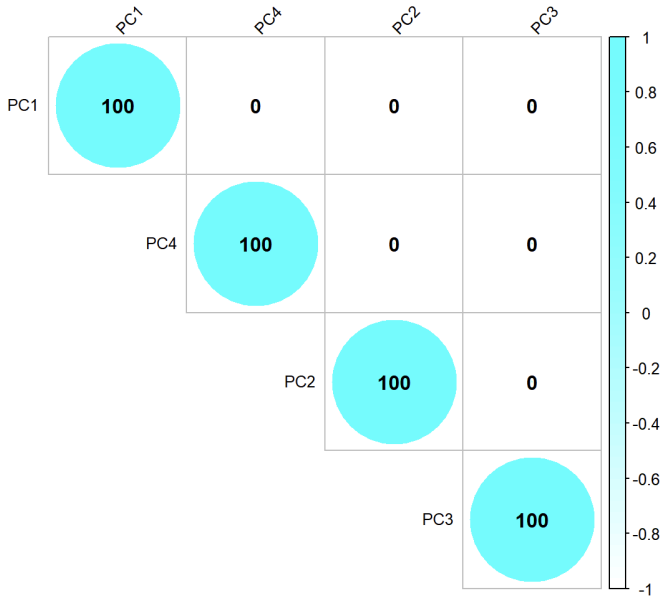

Figure 3.4. Correlation matrix for PCs (non-BIMARU)

Discussion and Analysis

Results of the multiple linear regression revealed statistical significance across the board. For BIMARU and non-BIMARU states, the principal components generated were statistically significant in terms of a linear relationship with every metric of economic growth. To better understand how these components explained the variation in the metrics of economic growth, we also looked at the R^2 values from the regression summaries.

The results of the analysis indicate that the infrastructure development variables have a significant impact on economic growth metrics in both regions, with varying degrees of magnitude. In the BIMARU states, the R^2 value for non-farm employment is 0.3466, indicating that **34.66%** of the variation in non-farm employment can be explained by the infrastructure development variables included in the model. Similarly, the R^2 value for agricultural employment is 0.4706, indicating that **47.06%** of the variation in agricultural employment can be explained by the infrastructure development variables.

In terms of consumption, the R^2 value is low in both regions, with a value of **0.02064** in the BIMARU states and **0.1347** in non-BIMARU states. This suggests that the infrastructure development variables have a relatively minor impact on consumption levels in both regions.

The R^2 value for PPP (poverty rate) is also relatively low in both regions, with a value of **0.0278** in the BIMARU states and **0.1248** in non-BIMARU states. This suggests that the infrastructure development variables have a limited impact on poverty levels in both regions.

Overall, the results of the multiple linear regression analysis suggest that infrastructure development variables have a significant impact on economic growth metrics across India, but the magnitude of the impact varies depending on the specific metric of economic growth being considered. In both regions, the infrastructure development variables have a greater impact on agricultural employment than on non-farm employment, and a relatively minor impact on consumption levels and poverty rates.

It is important to note that the R^2 values obtained from the multiple linear regression analysis do not necessarily indicate causality. While the results suggest that there is a relationship between infrastructure development variables and economic growth metrics, it is possible that other factors may be influencing these metrics. Therefore, further research is needed to fully understand the relationship between infrastructure development and economic growth in India.

As with any research, it is important to reflect on the limitations of our approach and acknowledge areas for future investigation. One notable limitation of our study was the high level of missing data, particularly in the BIMARU region, which led us to analyze a smaller number of SHRIDs than we initially intended. With our final analysis, we had ~190,000 data points for BIMARU and ~266,000 for the rest of India. While this limitation likely did not compromise the statistical significance of our findings, it may have reduced the generalizability of our results.

Another limitation of our study is the fact that we focused exclusively on rural areas, since SHRIDs are predominantly rural, as we discovered earlier in our research. Therefore, our findings may not necessarily extend to urban settlements, which may have different infrastructure development needs and economic growth patterns. Future research could explore whether the relationship between infrastructure development and economic growth differs between rural and urban areas, and how this might affect policy recommendations for improving economic outcomes in different regions. It would also be worthwhile to look at how these figures change across individual states, which could be used to advise state-specific policies for development. Additionally, the density of data that we are working with is only available for the last decade, owing to technological advancements. When data becomes available for upcoming years, it would help with longitudinal research and help us assess how policy changes can impact these metrics in the long run.

There is also scope for exploring social demographics as a potential factor. More specifically, the strength of the relationship between infrastructural development and economic growth could vary based on areas with lower/higher percentages of the population of marginalized groups.

While our analysis provides valuable insights into the relationship between X variables and Y variables using PCA and multiple linear regression, there are some limitations to our approach. Firstly, while we have the weights of each variable for each component derived from PCA, we did not extensively analyze the interpretability of these weights. The weights of the components can provide insights into which variables are most strongly related to the outcome variable, but without a thorough examination, the information derived from these weights may be limited. Understanding the weights of the components can help us identify important features that are driving the results, which can further inform our understanding of the relationships between variables.

Secondly, we did not thoroughly examine the implications of the weights of the regression associated with each principal component in X for each run of Y. This can be important for understanding how these aggregates fare when assessed in terms of explaining the variation in economic growth figures. The weights of the regression can also help us identify which variables are contributing to the outlier status, which can be useful in identifying and addressing potential problems in the data. But analyzing the weights of each component in the context of multiple linear regression with multiple Y variables can be challenging. As we ran multiple linear regression with multiple Y variables, there are many different combinations of weights to consider, which can be difficult to analyze and interpret. Therefore, we focused on the R^2 value to evaluate the model's performance, which provides a useful overall measure of the model's goodness of fit. However, as we have noted earlier, this approach has limitations, and examining the weights of the components can provide additional insights into the relationships between the variables. By relying solely on the R^2 value to analyze the model, we may have missed out on important nuances that could be explained by the weights of the components. This is because the R^2 value can sometimes lead to overfitting, where the model performs well on the training data but poorly on new, unseen data. By examining the weights of the components, we can get a better understanding of which variables are actually contributing to the model's performance and identify potential issues with overfitting.

References:

Rural population (% of total population) - India. World Bank. (n.d.). Retrieved March 13, 2023, from https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=IN

Contribution Of Rural Economy In New India Dr Vikash Singh Assistant Professor (Sahkaree P.G. Coolege Mihrawa ,Jaunpur) Veer Bahadur Singh Purvanchal University, Jaunpur, DOI: 10.47750/pnr.2022.13.S09.419

Ghosh, M. (2017). Infrastructure and development in rural India. *Margin: The Journal of Applied Economic Research, 11(3)*, 256–289. https://doi.org/10.1177/0973801017703499

M. S. Bhatia. (1999). Rural Infrastructure and Growth in Agriculture. *Economic and Political Weekly, 34(13)*, A43–A48. http://www.jstor.org/stable/4407793

Foster, A. D., & Rosenzweig, M. R. (2004). Agricultural Productivity Growth, Rural Economic Diversity, and Economic Reforms: India, 1970–2000. *Economic Development and Cultural Change, 52(3)*, 509–542. https://doi.org/10.1086/420968

Buddhadeb Ghosh, & Prabir De. (2004). How Do Different Categories of Infrastructure Affect Development? Evidence from Indian States. *Economic and Political Weekly, 39(42)*, 4645–4657. http://www.jstor.org/stable/4415682

Kaur, Avinash & Kaur, Rajinder. (2004). Role of Social and Economic Infrastructure in Economic Development of Punjab. *International Journal of Innovative Knowledge Concept. 5(6)*. 181-187. https://core.ac.uk/download/pdf/233155341.pdf

Ng, C. P., Law, T. H., Jakarni, F. M., & Kulanthayan, S. (2019). Road Infrastructure Development and Economic Growth. *IOP Conference Series: Materials Science and Engineering, 512*, 012045. https://doi.org/10.1088/1757-899x/512/1/012045

https://rural.nic.in/en/scheme-websites

Asher, S., Lunt, T., Matsuura, R., & Novosad, P. (2021). Development Research at High Geographic Resolution: An Analysis of Night Lights, Firms, and Poverty in India using the SHRUG Open Data Platform. The World Bank Economic Review.

Henderson, J. V., Storeygard, A., & Weil, D. N. (2011). A Bright Idea for Measuring Economic Growth. American Economic Review.