Refining State Level Comparisons in India

Pranjul Bhandari[1]
Planning Commission, Government of India

Working Paper Series, 2012

Abstract

In this paper we analyse the performance of Indian States across three critical sectors – health, education and infrastructure. To enable us to read through multiple indicators of the three sectors, we construct an index for each using the Principal Component Analysis technique. This technique assigns weights according to the relationship between the variables, thus involving relatively low levels of subjectivity on part of the researcher, while preserving most of the information in the original data set. Our 'raw' results conform with the already well-established findings of several other studies that states such as Kerala are amongst the best performing while the so-called BIMARU states (Bihar, MP, Rajasthan and UP) are laggards. While this is true on an absolute level, it does not reveal the performance conditional on state level factors. What we do next is refine this analysis. We control our three indices for per capita consumption to put the states on a level playing field and for gauging how well the states have used available resources. Our 'refined' analysis throws up rankings which are quite different from the 'raw' analysis. For instance, we find clear differentiation between the BIMARU states – while Orissa, Bihar and Chhattisgarh are amongst the best performers, Uttarakhand, Rajasthan and Jharkhand are amongst the worst. While the performance of Himachal Pradesh has been most impressive, Gujarat is amongst the worst on health, Maharashtra on infrastructure, and Haryana on both.

[1] I am grateful to Montek Singh Ahluwalia and Arunish Chawla from the Planning Commission, and Shriya Anand from the Indian Institute for Human Settlements for helpful comments and suggestions. Views and all errors are mine.

I. Introduction

The comparative performance of individual states has become an important area of research for a number of reasons. Given the well-known regional disparities in India, a study of parts (i.e. the states) becomes important if the sum of parts (i.e. the country) needs to progress in a balanced way. Also, a study of states throws up successful experiments and examples which can be replicated or adapted by other states. Issues at the state level are increasingly dictating election outcomes both at the centre and the states, making this study important for the political class as well. And finally, a comparative study can be useful for inducing some healthy competition across the states of India.

While Indian states can be compared across several criteria, in this paper we limit the comparison to three sectors – *health, education and infrastructure*. Each of these sectors is complex. Given the sheer size of resources needed for scale up, each of these three needs effort from both the public and private sectors. The public sector for instance not only needs to provide resources, but also create a policy environment conducive for scale-up.

In this paper, we try to analyse the long term performance of states in the provision of health and education services as well as infrastructure. We rank the states and gauge if performance across the three sectors are correlated or divergent. We compare states for both absolute performance as well as for performance after controlling for consumption levels. The latter analysis can be associated with governance - how well the resources at the state's disposal have been used for progress in the critical sectors of health, education and infrastructure. Our observations through the paper are limited to simple associations rather than causal relationships, which can be more complex to establish.

The rest of the paper is organised as follows: In section II, we construct separate indices for health, education and infrastructure across states. For each of the three sectors we combine a host of variables that are publicly available. We use the Principal Component Analysis technique to determine weights objectively. The three indices of health, education and infrastructure enable us to rank the states on their performance and also evaluate if good performance across the three are interlinked. We call this entire analysis a 'raw' comparison of states.

In section III, we refine the raw analysis of section II. It is well known over the last several decades that due to a variety of historic, social and economic reasons, while Kerala is a good performer in health and education outcomes, the so-called BIMARU states are laggards. What we do here instead is to control for per capita consumption before analysing or ranking performance. This puts the states on a level playing field before comparisons are made. For instance, Bihar's underperformance on many fronts could partly be explained by lower resources at its disposal which makes it difficult for the state to invest more on health and education. Our analysis controls for this factor while evaluating the state's performance in delivering key services. Figure 4 summarises our key findings.

In section IV, we compare the results from the raw and refined analysis. We conclude the paper with policy implications and scope for further research.

II. Raw comparison of States

In this section we draw comparisons across Indian States based on their progress on health, education and infrastructure. To make the comparisons easier to interpret, we make three separate indices (for health, education and infrastructure respectively), each of which combine several widely used and publicly available variables that are available across states. A description of the variables is given in **figure 1**. We cover 21 states in our analysis.

For health, we use both input (e.g. immunization) and output (e.g. Infant Mortality Rate) variables. For education, we use variables which reflect both the quantity (e.g. net enrolment rate) as well as quality (e.g. reading level for enrolled children). We break down infrastructure across sectors such as agriculture, electricity and transportation to ensure that the main sectors are included.

We use the Principal Component Analysis to assign weights to each of the variables. PCA becomes a useful variable reduction technique when the objective of the analysis is to present a huge data set using a fewer number of variables. It reduces the number of observed variables to a smaller number of principal components which account for

most of the variance in the observed variables[2]. PCA is used when the variables are highly correlated. If not, the analysis may be of no value. Of the various linear combinations, the first Principal Component, P1 (which we use here to calculate our composite index) is the one which accounts for the maximum possible proportion of the variance in the original dataset. The weights are termed as 'loadings' and depict how relevant the variable is in construction of the principle component. Because the weights are based on the relationships/correlations amongst the variables (caused by common 'factors'), this method involves relatively low levels of subjectivity on the part of the researcher.

[2] PCA decomposes a correlation matrix with ones (1s) on the diagonals. The amount of variance is equal to the sum of the diagonals (which is also the number of observed variables in the analysis) in the standardized dataset. Technically speaking, PCA minimizes the sum of the squared perpendicular distance to the axis of the principal component. The principal components account for a maximal amount of variance in the dataset. The component score is a linear combination of observed variables weighted by eigenvectors. If there are N variables - x_1, x_2, … x_n; P_1, P_2, … P_n are the N principal components, and a_{nn} are the weights, the first principal component can be written as a linear combination $P_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n$

Figure 1: Variables used for making the health, education and infrastructure indices

Variable	Source	Year
Health		
Life expectancy at birth (years)	Ministry of Health & Family Welfare	2006/10
Infant Mortality Rate (per 1,000 live births)	SRS	2010
Maternal Mortality Rate	SRS	2007/09
TFR (children per woman)	SRS	2009
Access to improved sanitation (%)	DLHS	2007/08
Proportion (%) of underweight children	NFHS	2005/06
Institutional Delivery (%)	DLHS	2007/08
Complete Immunization (%)	DLHS	2007/08
Education		
Mean years of schooling	NSS	2007/08
Female literacy rate, age 15+ years (%)	Census	2011
Aser - Reading level for enrolled children (Story)	ASER	2011
Aser - Arithmetic level for enrolled children (Division)	ASER	2011
Net Enrolment Ratio : Upper Primary Level	HRD	2009/10
Dropout rate (I-VIII)	HRD	2009/10
Infrastructure		
Agriculture:		
Gross irrigated area/gross cultivated area	Ministry of Agriculture	2008/09
Communication:		
Teledensity/1000 population	Department of Telecommunications	2008/09
Post Offices/1000 population		2007/08
Banking:		
Bank branches/1000 population	RBI	2008/09
Electricity:		
Electricity consumption/1000 population	Central Electricity Authority	2008/09
% of villages electrified		2008/09
Installed capacity/1000 population		2008/09
Length of T&D lines/1000sq km		2008/09
Transportation:		
Total surfaced highways/1000 sq km	Ministry of Road Transport and Highways	2007/08
Other surfaced roads/1000 sq km		2007/08
Registered motor vehicles in 1000s/1000 sq km		2008/09
Railroad length/1000 sq km		2007/08

The methodology entails the following steps – first, we get a complete data set of all the variables across the 21 states. We order the data such that 'higher is better'. For example, higher institutional deliveries are better and the data is left as is. But higher Infant Mortality Rate is worse, therefore we take the inverse of IMR. Since variables measured at different scales do not contribute equally to the analysis, we standardise the data set (by subtracting the mean value of each variable across states and dividing by its standard deviation). Now each variable has a mean of zero and a standard deviation of 1. Finally, we apply the PCA analysis on this standardised dataset in order to calculate the weights and form the weighted index. In our analysis, no negative

weights have been observed. Since our dataset is standardised, each of the three indices have a zero mean. The Principal Component for our three indices explains 60 – 80% of the variation among the variables.

While the health and education indices involve one round of principal component analysis, we use a two stage PCA technique for infrastructure. There are various sub-sectors for infrastructure, several of which have more than one variable. We fist use the PCA analysis to get an index each for the sub sectors which have more than one variable. We then apply PCA again to the subsectors to get the final infrastructure index.

We rank the three indices in **Figure 2**. For ease of illustration, we eyeball the rankings and put them in 3 tiers of seven states each. The following points stand out –

- The **first tier states** comprising Kerala, Goa, Himachal, Punjab, Tamil Nadu, Maharashtra and Haryana are the best performers. However, performance of Maharashtra in infrastructure and that of Haryana in health is markedly poor.
- The **second tier states** comprising West Bengal, Uttarakhand, Karnataka, Andhra, Gujarat, J&K and Orissa are the medium performers. Orissa stands out for worse performance on infrastructure, compared to its performance in health and education.
- The **third tier states** comprising Rajasthan, Assam, MP, Chattisgarh, UP, Bihar and Jharkhand are the laggards, mostly comprising of the BIMARU states.

The rank correlation between the three indices is high, ranging from 81% to 88%, implying similarities in performance across health, education and infrastructure. Of the three correlations, the one between health and education is the highest. The rank correlation between each of the three indices and monthly per capita consumption expenditure (MPCE; source: NSSO, 2009/10) is also high, ranging between 80% and 87%. While these are simple associations and not causal relations, they suggest that higher growth and income are associated with better health, education and infrastructure status.

Figure 2: Three tiers in ranking Health Education and Infrastructure

	Ranks across States	Health Index Ranks	Education Index Ranks	Infrastructure Index Ranks
First Tier	Kerala	1	1	3
	Goa	2	3	1
	Himachal	6	2	2
	Punjab	4	6	4
	TN	3	8	5
	MH	5	4	11
	Haryana	11	5	9
Second Tier	West Bengal	7	9	12
	Utt	13	7	7
	Karnataka	9	11	8
	Andhra Pradesh	8	12	10
	Gujarat	12	10	6
	J&K	10	15	14
	Orissa	14	14	17
Third Tier	Rajasthan	15	16	15
	Assam	16	13	19
	MP	20	18	13
	Chtts	17	17	18
	UP	21	21	16
	Bihar	19	19	20
	Jharkhand	18	20	21

Rank correlation bw -			
Health and Education	**0.88**	Health andMPCE	**0.80**
Education and Infrastructure	**0.85**	Education and MPCE	**0.86**
Infrastructure and Health	**0.81**	Infrastructure and MPCE	**0.87**

III. Refined comparison of States

While the analysis above is insightful, it only reiterates the well known fact that states like Kerala have done well on health and education, while the BIMARU states have been laggards. States with lower resources at their disposal are likely to underperform. In this section, we refine our analysis by creating a level playing field before comparing states. We adjust the three indices created in section 1 for monthly per capita consumption (MPCE).

Although GDP per capita and consumption per capita broadly measure the same thing and are tightly correlated (with a correlation coefficient of 90%), consumption has the benefits of reflecting the actual purchasing power and including income generated from outside the state (i.e. inter state remittances). We calculate state wise MPCE by taking a population weighted average of rural and urban MPCE for each state.

Population statistics are taken from the Census 2011, and rural and urban MPCE from NSSO 2009/10. To control for MPCE, we run semi-log OLS regressions between the three indices and MPCE –

HEALTH = -19.02 + 2.63 * log (MPCE)
t stat = 7.34, R-squared = 0.74

EDU = -16.08 + 2.22 * log (MPCE)
t stat = 6.03, R-squared = 0.66

INFRA = -14.64 + 2.02 * log (MPCE)
t stat = 6.68, R-squared = 0.70

In each of the three regressions, the coefficients are significant at the 1% level. The R-squared ranges between 66% and 74% suggesting a good fit. We also run the regressions with the log of per capita GDP instead of MPCE, but while the coefficients remain significant, the R-squared lowers (to the 57 – 66% range)[3]. As shown in **figures 3a, 3b and 3c**, the regression gives us the line of best fit across the 21 states of India. The positive slope highlights the long term positive and highly significant association between consumption and the three indices - health, education and infrastructure.

What the regressions also throw up are the residuals. Positive residuals (i.e. states lying above the line of best fit) are better than what the average all-India performance suggests, and negative residuals (i.e. states lying below the line of best fit) are worse than what the average all-India performance suggest.

[3] MPCE works well for health and education as both are household decisions to a large extent. While it could be argued that GDP per capita should be used for infrastructure, we continue to use MPCE because (a) R squared is better with MPCE and (2) using MPCE for each of the three sectors is important for doing a comparable analysis.

Figure 3a: The good and bad performers in health

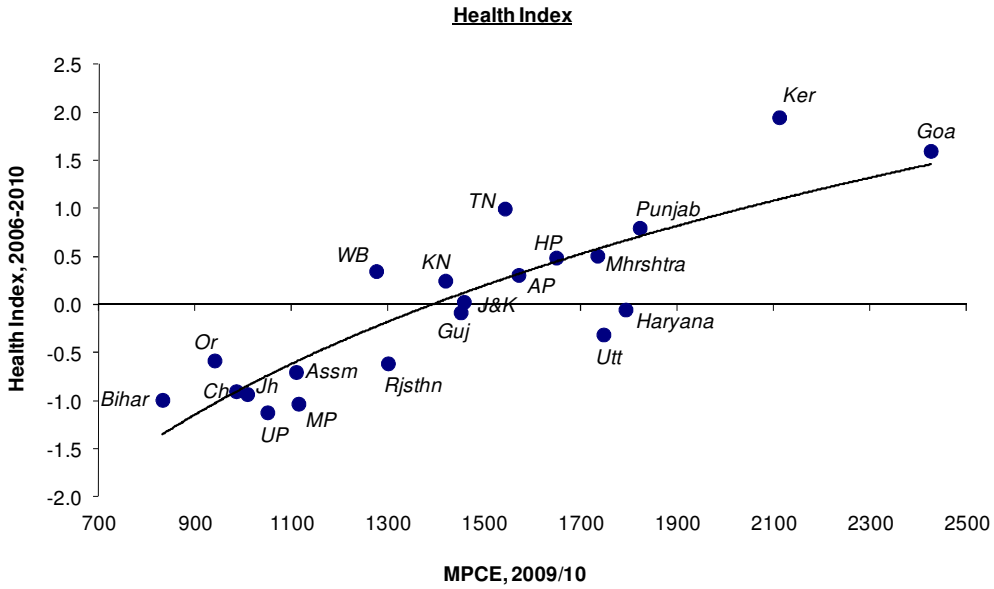


Figure 3b: The good and bad performers in education

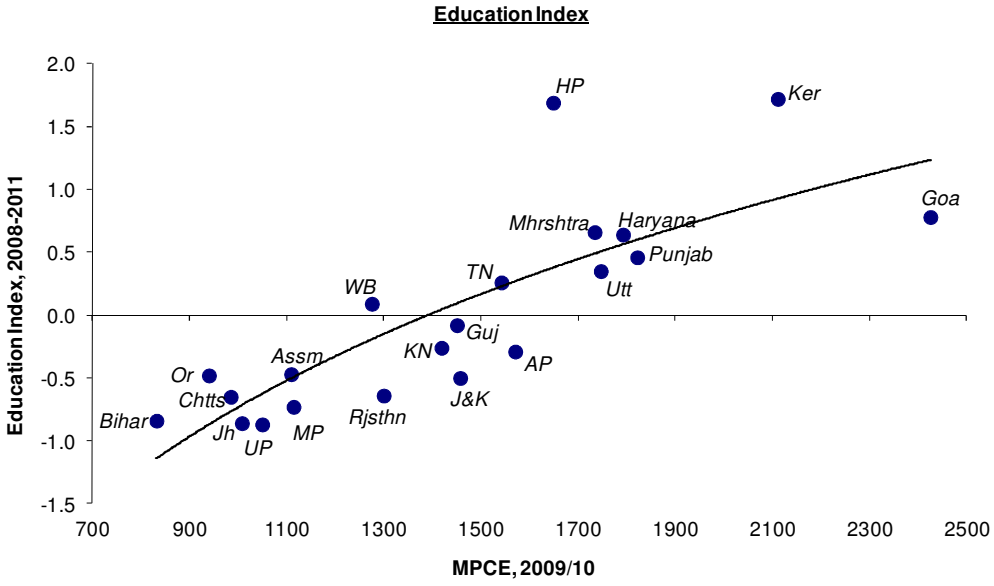
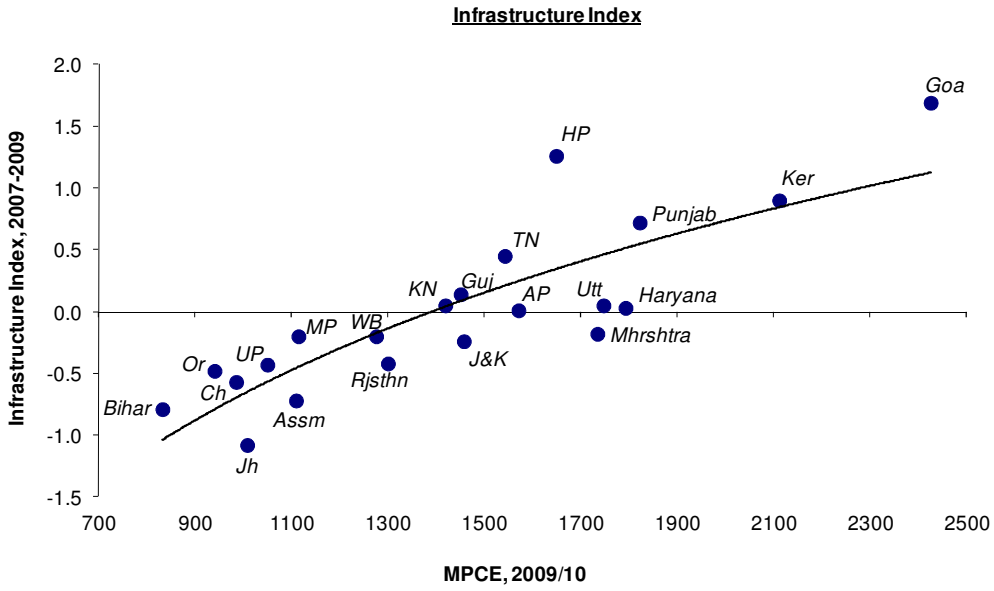


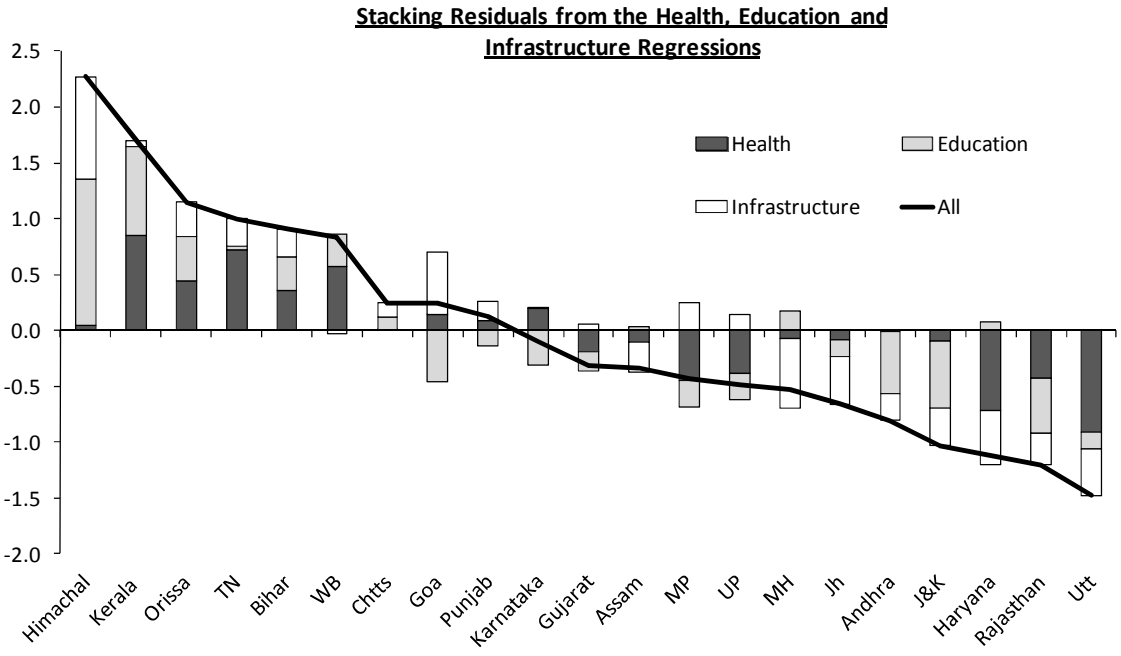
Figure 3c: The good and bad performers in infrastructure



We stack up the residuals from the three regressions in **figure 4.** The 'refined' analysis throws up the following observations -

- **Good performers - Himachal Pradesh, Kerala, Orissa, Tamil Nadu and Bihar** have been the best performers across all the three sectors. **West Bengal and Chattisgarh** have also been amongst the best off states.
- **Laggards - Uttarakhand, Rajasthan, J&K and Jharkhand** have been laggards across all the three sectors.
- **Average performers -** The remaining middle ranking states have varied performance. **Goa, Punjab and Karnataka** have done well in health and infrastructure, but underperformed in education. On the other hand, **Haryana, Andhra, Gujarat, Assam, MP, UP and Maharashtra** have each underperformed in two of the three sectors we have analysed.

Figure 4: Stacking up performance across States



We also rank the states across health, education and infrastructure based on the residuals. The rank correlations between them have fallen to the 25% to 50% range (*46% between health and education; 25% between education and infrastructure; 50% between infrastructure and health*) compared to the 80% to 87% range in the raw analysis. This was expected given that we have now controlled for consumption which could have been directly or indirectly driving some of the similarities in rankings in the raw analysis of section II.

IV. Comparing raw and refined analysis of States

As shown in **figure 5**, the rankings of many states change when the indices are refined -

- **Bihar, Orissa and Chattisgarh** have risen sharply in rankings across all the three sectors. Relative ranking of **Jharkhand** has also improved but it remains a laggard state.

- **Haryana and Uttarakhand** have fallen in rankings across all the three sectors**. Gujarat, Punjab and Maharashtra** have also slipped in ranks in the refined analysis.

Figure 5: Raw vs. refined rankings of States

	HEATH			EDUCATION			INFRASTRUCTURE		
		Refined ranks	Raw ranks		Refined ranks	Raw ranks		Refined ranks	Raw ranks
First Tier	Kerala	1	1	HP	1	2	HP	1	2
	TN	2	3	Kerala	2	1	Goa	2	1
	WB	3	7	Orissa	3	14	Orissa	3	17
	Orissa	4	14	Bihar	4	19	Bihar	4	20
	Bihar	5	19	WB	5	9	MP	5	13
	Karnataka	6	9	MH	6	4	TN	6	5
	Goa	7	2	Chtts	7	17	Punjab	7	4
Second Tier	Punjab	8	4	Haryana	8	5	UP	8	16
	HP	9	6	Assam	9	13	Chtts	9	18
	Chtts	10	17	TN	10	8	Kerala	10	3
	Andhra	11	8	Punjab	11	6	Gujarat	11	6
	MH	12	5	Jharkhand	12	20	Karnataka	12	8
	Jharkhand	13	18	Utt	13	7	WB	13	12
	J&K	14	10	Gujarat	14	10	Andhra	14	10
Third Tier	Assam	15	16	MP	15	18	Assam	15	19
	Gujarat	16	12	UP	16	21	Rajasthan	16	15
	UP	17	21	Karnataka	17	11	J&K	17	14
	Rajasthan	18	15	Goa	18	3	Utt	18	7
	MP	19	20	Rajasthan	19	16	Jharkhand	19	21
	Haryana	20	11	Andhra	20	12	Haryana	20	9
	Utt	21	13	J&K	21	15	MH	21	11

Rank tier rises after refining
Rank tier falls after refining

V. Conclusion

There is enormous scope of further research in analysing the performance of states. The 'refined' analysis can be conducted every few years to monitor incremental changes, or the regression could be run on growth rather than levels over specified time periods. This will allow us to gauge how particular states are improving their performance over time and how performance across different time periods has differed. While we have controlled for consumption, other variables or combination of variables which cover economic, social, biological, etc differences across states can also be used.

The refined analysis of states throws up important results on which states are making best use of the resources in hand to provide health, education and infrastructure services to its people. It is therefore a useful tool in identifying states whose experiments are working, and which can potentially be replicated by others. While convergence in income levels may take its own time, this analysis will help policy experts, interested observers and even voters to evaluate the success of its state and government.