A Method of Classifying Regions from Multivariate Data

N S Iyengar

P Sudarshan

In regional studies, it is customary to use a composite index to measure development. In this note an attempt is made to define a composite index for measuring the spatial differentials in the level of development.

An illustration is provided for the construction and use of such indices, taking the available district-wise data from the states of Andhra Pradesh and Karnataka.

I

Measuring Distances among Regions

A NUMBER of studies have been carried out to identify backward regions using various criteria of development. Bennet [2], for example, constructed a non-monetary index of development to focus attention on international disparities in consumption levels. Adelman *et al* [1] refined Bennet's analysis by incorporating additional variables. In India, Rao [15] used multiple factor analysis approach for measuring economic distance between the states. Das Gupta [3] considered some 24 indicators used earlier by Mitra [7] for classifying the various districts of India on a ranking basis, and used the discriminant analysis. Rao [14], taking 15 indicators and using the method of principal components, identified the backward states.

In all these studies, generally, a factor analysis approach was adopted. Nanjappa [12] used an approach involving a simple aggregation of the rankings of districts in Karnataka state, based on some 15 development indicators. The Pande Committee [13] adopted a similar ranking approach at the all-India level. Mukherjee and Roy [10,11] have recently proposed a more sophisticated approach analogous to factor-analysis. In a somewhat unconventional approach, Hellwig [4] had constructed an index of development in the form of a weighted average, where the weights are assumed to be inversely proportional to the coefficients of variation. This index was used for classification of countries. Iyengar *et al* [5] independently developed a similar index for measuring development of districts in Karnataka. Sudarshan [16] has proposed yet another method for classifying districts in Andhra Pradesh.

In this paper, we emphasise the spatial aspects of development by proposing a simple method[1] for measuring the level or stage of district development. The methodology is described in Section II. A practical application of this method is considered in Section III by using a selected number of development indicators for Andhra Pradesh and Karnataka. It will be seen that this method is a simpler and probably a better alternative to the conventional approaches, such as the principal component analysis, which are based on rather restrictive assumptions.[2]

II

Methodology

Let X_{id} represent the size or value of the i-th development indicator in the d-th district of a state (i=1,2,..., m: d = 1, 2,...n, say.) Let us write

$$y_{id} = \frac{X_{id} - \min_d X_{id}}{\max_d X_{id} - \min_d X_{id}} \quad .. (2.1)$$

where $\min_d X_{id}$ and $\max_d X_{id}$ are, respectively, the minimum and maximum of $(X_{i1}, X_{i2}, \ldots, X_{in})$.

If, however, X_i is negatively associated with development, as, for ex-

TABLE 1: VARIABLE INDICATORS AND THEIR WEIGHTAGE

	Indicator	Weight
Education		
X_1	Literacy rate	.0426
Health		
X_2	Number of hospital beds per lakh of population	.0515
X_3	Doctors per lakh of population	.0540
Agriculture		
X_4	Percentage of area irrigated to the area sown	.0448
X_5	Yield per hectare	.0467
X_6	Percentage area under commercial crops to the area sown	.0430
X_7	Number of pumpsets	.0452
X_8	Number of tractors	.0429
Industry		
X_9	Value-added by manufacturing	.0525
X_{10}	Factory employment	.0469
Infrastructure		
X_{11}	Percentage of villages connected by all-weather roads	.0430
X_{12}	Kilometerage of surfaced roads per 1000 square kilometre area	.0434
X_{13}	Motor cars and jeeps	.0543
X_{14}	Motor cycles and scooters	.0540
X_{15}	Goods vehicles on road	.0482
X_{16}	Radio sets	.0447
X_{17}	Telephones	.0538
X_{18}	Number of post offices per lakh of population	.0484
Other infrastructurals		
X_{19}	Percentage of villages and towns electrified	.0446
X_{20}	Per capita consumption of electricity	.0451
X_{21}	Bank offices (scheduled) per lakh of population	.0502

TABLE 2: LEVEL OF DEVELOPMENT IN ANDHRA PRADESH DISTRICTS (1978-79)

District	Index of Level of Development (%)	Rank
Coastal Andhra		
East Godavari	35.42	6
Guntur	37.51	4
Krishna	41.59	3
Nellore	30.10	8
Prakasham	21.22	13
Srikakulam	14.78	19
Vishakapatnam	33.01	7
West Godavari	44.08	2
Rayalaseema		
Anantapur	26.51	10
Chittur	37.49	5
Cuddapah	26.29	11
Kurnool	27.78	9
Telangana		
Adilabad	10.83	21
Hyderabad	68.38	1
Karimnagar	20.25	15
Khammam	17.68	17
Mahboobnagar	14.17	20
Medak	14.86	18
Nalgonda	20.93	14
Nizamabad	23.39	12
Warangal	19.99	16

TABLE 3: CLASSIFICATION OF ANDHRA PRADESH DISTRICTS (1978-79)

Stage of Development	District
Highly Developed	Hyderabad West Godavari Krishna
Developed	Guntur Chittur East Godavari Vishakapatnam Nellore
Developing	Kurnool Anantapur Cuddapah Nizamabad
Backward	Prakasham Nalgonda Karimnagar Warangal Khammam
Very Backward	Medak Srikakulam Mahboobnagar Adilabad

ample, the infant mortality rate or the unemployment rate which should decline as the district develops, then (2.1) can be written as:

$$y_{id} = \frac{\max_d X_{id} - X_{id}}{\max_d X_{id} - \min_d X_{id}} \qquad (2.2)$$

Obviously, the scaled values, y_{id}, vary from zero to one[3]. From the matrix of scaled values, $Y = ((y_{id}))$, we may construct a measure for the level or stage of development for different districts as follows:

$$\bar{y}_d = w_1 y_{1d} + w_2 y_{2d} + \dots + w_m y_{md} \qquad (2.3)$$

where the w's ($0 < w_i < 1$ and $w_1 + w_2 + \dots w_m = 1$) are arbitrary weights reflecting the relative importance of the individual indicators. A special case of this is when the weights are assumed equal.

However, a more rational view would be to assume that the weights vary inversely as the variation in the respective indicators of development. More specifically, we shall assume:

$$w_i = \frac{k}{\sqrt{\operatorname{Var}(y_i)}} \quad \dots (2.4)$$

$$\text{where} \quad k = \left[\sum_{i=1}^{m} \frac{1}{\sqrt{\operatorname{Var}(y_i)}} \right]^{-1} \quad \dots (2.5)$$

The overall district index, $\bar{y}_d$, also varies from zero to one. Also, if $y_1, y_2, \dots, y_m$ are independent, then

$$\operatorname{Var}(\bar{y}_d) = \sum_{i=1}^{m} w_i^2 \operatorname{Var}(y_i) \quad \dots (2.6)$$

which is constant, equals to mk^2 for all the districts.

The choice of the weights in this manner ensures that large variation in any one of the indicators will not unduly dominate the contribution of the rest of the indicators and distort interdistrict comparisons. It is well-known that, in statistical comparisons, it is more efficient to compare two or more means *after* equalising their variances.

For classificatory purposes, a simple ranking of the district indices ($\bar{y}_d$) would do. However, a more meaningful characterisation of the different stages of development would be in terms of suitable fractile classification from an assumed distribution of $\bar{y}$. It appears appropriate to assume that $\bar{y}$ has a Beta distribution in the range (0,1). The Beta distribution is generally skewed, and perhaps, relevant to characterise positive-valued random variables.

A random variable, Z, has a Beta distribution in the interval (0,1) if its probability density function, f(z), can be written as:

$$f(z) = \frac{1}{B(a,b)} z^{a-1} (1-z)^{b-1},$$

$$0 < z < 1 \text{ and } a, b > 0 \quad \dots (2.7)$$

where B(a,b) is the integral

$$B(a,b) = \int_0^1 z^{a-1} (1-z)^{b-1} dz \quad \dots (2.8)$$

Let $(0, z_1)$, (z_1, z_2), (z_2, z_3), (z_3, z_4); and $(z_4, 1)$ be linear intervals, such that each interval has the same probability weight of 20 per cent. These fractile groups can be used to characterise the various stages of development. Suppose we adopt the following definitions of development, excluding the extreme cases of z = 0, 1.

E Very backward if

$0 < \bar{y}_d \leq z_1$

D Backward if

$z_1 < \bar{y}_d \leq z_2$

C Developing if

$z_2 < \bar{y}_d \leq z_3$

B Developed if

$z_3 < \bar{y}_d \leq z_4$

A Highly developed if

$z_4 < \bar{y}_d < 1$.

The parameters (a, b) in the assumed Beta distribution can be estimated by solving the following simultaneous equations:

$$(1-\bar{y})a - \bar{y}b = 0$$

$$(\bar{y} - m_2)a - m_2 b = m_2 - \bar{y} \quad \dots (2.9)$$

where $\bar{y}$ is the overall mean of the district indices and m_2 is given by

$$m_2 = S_y^2 + \bar{y}^2 \quad \dots (2.10)$$

where S_y^2 is the variance of the district indices. The cut-off points z_1 to z_4 can be obtained from tables of incomplete Beta function, or from table of the F-distribution with degrees of freedom (2a, 2b), which are readily available.

If $F_{n_1, n_2; p}$ is the value of the F-Statistic with n_1 and n_2 degrees of

TABLE 4: INDICATORS AND THEIR WEIGHTAGE

Indicator	Weight
Education and culture	
Literacy rate	0.04462
Cinema houses/lakh of population	0.04674
Health	
Hospital beds/lakh of population	0.05571
Agriculture	
Percentage of gross area irrigated/gross aea sown	0.05325
Percentage of area under commercial crops to the gross area sown	0.05472
Number of water pumps	0.04706
Number of tractors	0.05700
Percentage of area under high yielding varieties	0.04635
Industry	
Factory employment	0.06276
Infrastructure	
Length of surfaced roads/1000 sq km area	0.06736
Motor cycles & scooters	0.06191
Number of cars & jeeps	0.06227
Goods vehicles	0.05911
Radiosets	0.06131
Telephones	0.06219
Post offices/lakh of population	0.05929
Percentage of towns & villages electrified	0.04885
Bank offices/lakh of population	0.04949

freedom corresponding to probability p, i e,

$$\Pr (F \leq F_{n_1, n_2; p}) = p \quad \ldots (2.11)$$

then

$$F_{n_1, n_2; p} = \frac{n_2}{n_1} \frac{1 - z_p}{z_p} \quad \ldots (2.12)$$

where z_p is the p^{th} fractile of the corresponding Beta distribution.

Hence in our case, z_p is given by

$$z_p = \frac{1}{1 + \frac{b}{a} F_{n_2, n_1; p}} \quad \ldots (2.13)$$

since $n_1 = 2a$, $n_2 = 2b$. Extensive tables are available for computing the fractile points on the F-distribution for selected values of (n_2, n_1) and p. For values of F not readily available in the tables a two-way interpolation is needed. A straightforward procedure would be as follows:

For values of p less than 0.5, let $F_{n_{2k}, n_{1k}}$ be the tabulated value of the F-ratio with degrees of freedom (n_{2k}, n_{1k}) for a given fracticle point on the F-distribution. Taking $k = 1$ and $k = 2$, we wish to compute, say, F_{n_2, n_1} for values of (n_2, n_1) where $n_{21} < n_2 < n_{22}$ and $n_{11} < n_1 < n_{12}$. It is easy to show that

$$F_{n_2, n_1} = F_{n_{21}, n_{11}} + \frac{n_2 - n_{21}}{n_{22} - n_{21}} (F_{n_{22}, n_{11}} - F_{n_{21}, n_{11}}) + \frac{n_1 - n_{11}}{n_{12} - n_{11}} (F_{n_{21}, n_{12}} - F_{n_{21}, n_{11}}) + \frac{(n_2 - n_{21})\,(n_1 - n_{11})}{(n_{22} - n_{21})\,(n_{12} - n_{11})} \left[F_{n_{21}, n_{11}} + F_{n_{22}, n_{12}} - F_{n_{21}, n_{12}} - F_{n_{22}, n_{11}} \right]$$

However, for $p \geq 0.5$ the following result holds:

$$F_{n_1, n_2, p} = \frac{1}{F_{n_1, n_2, 1-p}}$$

III

Illustration: Andhra Pradesh

To illustrate the method described above, we have used the district-wise data for Andhra Pradesh, taken from Sudarshan [16], which involves 21 districts and 21 individual indicators of development. The basic data relate to 1978-79, and were obtained from published and unpublished sources of official statistics readily available. The estimated weights of the scaled variables are shown in Table 1.

The indices of development are presented in Table 2 for all the districts considered, along with their relative rankings.

These indices are further graduated using the Beta distribution with the estimated parameters, a = 2.9087 and b = 7.5094. The 20 per cent cut-off points are estimated to be: 0.1600, 0.2303, 0.3009, 0.3913. Based on these calculations, the Andhra Pradesh districts are classified into five clusters according to their stage of development, as shown in Table 3.

According to our exercise, the entire Telangana region of Andhra Pradesh is backward — except the capital district of Hyderabad and Nizamabad. In the coastal Andhra region, Srikakulam and Prakasham districts are 'very backward' and 'backward', respectively. Two districts of Andhra region are highly developed, and the rest are developed. In the Rayalaseema region, it turns out that Chittur is the most developed and the remaining three districts are developing.

IV

Illustration: Karnataka

As a second illustration, we examine the spatial differentials in Karnataka's development. Karnataka consists of 19 districts. We have considered 18 broad indicators of development. The choice of indicators was dictated by the ready availability of secondary data at the Karnataka Bureau of Economics and Statistics. These data relate to 1980-81. The indicators used and their respective weights are shown in Table 4.

In Table 5 we give the stage of development index for each of the 19 districts of Karnataka, followign the same methodology as was used in the case of Andhra Pradesh.

These indices were further graduated, using a continuous Beta distribution of the first type, with estimated parameters of a = 3.250156 and b = 8.33645. The 20 per cent cut-off points were found to be 0.1673, 0.2344, 0.3019, and 0.3874. Based on these calculations, the Karnataka districts were finally classified into five clusters according to their stage of development, as shown in Table 6.

TABLE 5: LEVEL OF DEVELOPMENT IN KARNATAKA DISTRICT (1980-81)

District	Index of Level of Development (%)	Rank
Bangalore	69.98	1
Dakshina Kannada	42.20	2
Kodagu	33.99	3
Shimoga	32.38	4
Dharwad	32.25	5
Belgaum	31.00	6
Bellary	28.06	7
Mandya	27.84	8
Chickmaglur	26.93	9
Mysore	26.25	10
Chitradurga	26.17	11
Uttara Kannada	26.08	12
Kolar	26.05	13
Hassan	23.36	14
Tumkur	20.89	15
Bijapur	20.40	16
Raichur	16.50	17
Bidar	12.79	18
Gu'barga	9.84	19

TABLE 6: CLASSIFICATION OF KARNATAKA DISTRICTS (1980-81)

Stage of Development	District
Highly developed	Bangalore
	Dakshna Kannada
Developed	Kodagu
	Shimoga
	Dharwad
	Belgaum
Developing	Bellary
	Mandya
	Chickmagular
	Mysore
	Chitradurga
	Uttara Kannada
	Kolar
Backward	Hassan
	Tumkur
	Bijapur
Very backward	Raichur
	Bidar
	Gulbarga

Our exercise indicates that the entire Hyderabad-Karnataka region, *viz*, Raichur, Bidar and Gulbarga, are still 'very backward', while Bijapur is slightly above these three districts. Hassan is very close to the 'developing' stage, while Tumkur, surprisingly, is still 'backward'. In the developing category, there are as many as seven districts exhibiting a good measure of homogeneity. The developed districts are, Kodagu, Shimoga, Dharwad, Belgaum, Dakshina Kannada, and Bangalore, of which the last two are 'highly developed'. Bangalore district has the distinction of being the best developed district, while Gulbarga remains at the lowest position in the scale of development.

V

Concluding Remarks

In the list of indicators used in our illustrations some very important and highly relevent indicators, such as the per capita calorie intake, per capita consumption of proteins per day, per capita consumption of cloth, life expectancy, infant mortality rates, etc. are not included. This is mainly because of non-availability of data at the ditsrict level. Also, in our list, one finds the dominance of the infrastructural indicators. It may be pointed out that this choice was deliberate, since it is well recognised that infrastructural development is a necessary precondition for rapid development and pomotion of social justice. The remarkable changes that have taken place in recent years in Taiwan and Korea serve as good examples of this.

One might also argue that some of the indicators employed in this study are superfluous, but this argument does not hold water when one recognises, for example, that the percentage[7] of villages electrified reflects the rural development aspect, whereas the per capita consumption of electricity reflects some other aspects of development. Development being a complex multi-dimensional phenomenon, one cannot altogether avoid using different indicators simultaneously, which may appear redundant at first sight and which may in fact be not quite so — as the above-mentioned example shows.

Any index of development based on multivariate data has its own limitations. A major limitation arises from the assumptions made about the indicators themselves and their weightage in the aggregate index: We believe that any inter-district comparison of levels of development would be more efficient when the variability in the composite index is stabilised. In the special case, when there are only two districts to be compared, the indicators will assume only two values, 0 and 1; and each indicator will hence have the same variance of one-half. This result, however, does not apply to cases involving more than two districts. Taking the example of three districts, we find that the weightages are equal only if one of the three districts lies exactly half way between the other two districts.

However, in the two illustrations we have considered, the distribution of weights among various indicators appears more or less uniform. It is also found that the clustering of the districts is not unduly affected by assigning equal weightage. One possible explanation for this can be that the original variables (X) are already weighted once, by using the respective ranges as a measure of variability in arriving at the scaled variables (Y). Thus, it appears that, for all practical purposes, it does not matter whether one uses a weighted average or a simple average of the scaled values for constructing the composite index.

Another methodological question could be regarding our adoption of the Beta distribution for graduating the district indices. Here again, we were guided by pragmatic considerations. Graduation using a normal distribution could have been resorted to, but the Beta distribution was preferred because of its skewness and its finite range. And these are precisely the properties to look for in statistical models suitable for analysing economic size distributions. The Chi-square test of goodness of fit, in both the illustrations, also confirms that the Beta distribution is more appropriate. The critical values of the Chi-square statistic at 5 per cent and 1 per cent level of significance for two degrees of freedom are, respectively, 5.99 and 9.21, which far exceed the computed values of the Chi-square statistic.

It should be pointed out that, in our analysis, we do not regard any district as fixed for purposes of comparison. The determination of such standard district or norm would be statistically and conceptually very difficult. Also, certain indicators in our exercise may not be spatially comparable since the district sizes are unequal.

In spite of the limitations discussed above, our analysis brings out in quantitative terms certain interesting aspects of development in the states of Andhra Pradesh and Karnataka. Indian planners may find our methodology particularly attractive and useful in their regional analyses for arriving at rational decisions on allocation of resources to develop the backward areas.

Notes

[The authors are grateful to H N Nagaraja for his comments on an earlier draft of this paper. They are also grateful to D Dharmappa for his excellent typing of the paper.]

1 This method is analogous to the one proposed by Morris and Liser [8], and used by Mukherjee [9] for inter-state comparison.

2 The principal component analysis assumes that the variable indicators are linearly related. When non-linearity is present, the component analysis is not appropriate. Further, one cannot assign any specific economic meaning to the transformed variables. They are artificial orthogonal variables not directly identifiable with a particular economic magnitude. See, for example Koutsoyiannis (6, p 436).

3 This transformation may appear similar to the practice of measuring the deviations from the mean in units of standard deviation, often resorted to in applied statistical work in areas like psychology. But the latter practice has certain disadvantages as far as the interpretation is concerned. On the other hand, the transformation employed here has a natural meaning in the context of measurement of development, which is always a relative concept.

References

[1] Adelman, Irma; Papelasis; and Lean Mears: Economic Development Analysis: Case Studies", Universal Book Stall, New Delhi, 1967.

[2] Bennet, M K: 'International Disparities in Consumption Levels', *American Economic Review*, XLI (Sepember 1951), pp 632-649.

[3] Das Gupta B: 'Socio-economic Classification of Districts: A Statistical Approach', *Economic and Political Weekly*, August 14, 1971.

[4] Hellwig, Z: 'On the Problem of Weighing in International Comparisons', in "Towards a System of Human Resources for Less Developed Countries" (*Ed* by Gostkowski), Institute of Philosophy and Sociology, Polish Academy of Sciences.

[5] Iyengar, N S; Nanjappa, M B; and Sudarshan, P: 'A Note on Inter-District Differentials in Karnataka's Development' (To be published in the *Journal of Income and Wealth*, Volume 5, No 2, 1981).

[6] Koutsoliannis: "Theory of Econometrics" 2nd Edition 1977.

[7] Misra, Ashok: 'Levels of Regional Development in India', Census of India 1961, Part I A (i).

[8] Morris, M D and Liser, P B: 'The PQLI: Measuring Progress in Meeting Human Needs' Overseas Development Council, *Communique on Development Issues*; No 32, 1977.

[9] Mukherjee, M: "Physical Quality of Life Index", Centre for Monitoring Indian Economy, Bombay, 1980.

[10] Mukherjee, M. and Roy, A K, 'A Method of Combining Diverse Partial Measures of Development', *The Journal of Income and Wealth*, Volume 2, No 1, 1977, pp 48-51.

[11] —. —: 'A Class of Methods of Combining Diverse Partial Measures of Development', *The Journal of Income and Wealth*, Volume 3, No 1, 1978, pp 36-38.

[12] Nanjappa, M B: "Backward Areas in Mysore State: A Study in Regional Development", *Southern Economist*, 1968.

[13] The Pande Committee Report: "Identification of Backward Regions" Government of India, 1968.

[14] Rao, Hemalatha: 'Identification of Backward Regions and the Trends in Regional Disparities in India', *Artha Vignana*, Volume 9 No 2, Nov 1977, pp 93-112.

[15] Rao, S K: 'A Note on Measuring Economic Distance between Regions in India', *Economic and Political Weekly*, April 28, 1973.

[16] Sudarshan, P: "Spatial and Inter-Temporal Aspects of Development in Andhra Pradesh". (M Phil Thesis submitted to Kakatiya University, 1981).