Trinary Clustering Framework for Modeling Multiple Parameters and Cluster-Based Interactions

Turkan K. Gardenier and John S. Gardenier
Pragmatica Corp., 1733 Kirby Rd., Unit 1408, McLean, VA 22101
National Center for Health Statistics (Retired) 1733 Kirby Rd., Unit 1408, McLean, VA

Abstract

A Trinary or Trinomial based orientation to Big Data analytics has the advantage of identifying relevant regions within large datasets for further exploration. This approach also formulates subsets or clusters which can be linked to multiple parameters. The Trinary (-/0/+) regions inherent in nominal or ordinal based distributions e3ssentially characterize input records feasible for multi-parameter linking. The authors have demonstrated applications to air quality monitoring and Geographic Information Science (GIS) based depiction of lung cancer mortality rates. This paper deals with applications to international indicators such as ethnic and linguistic fractionalization, GINI index, GDP per capita USD, and Happiness Index derived from the GALLUP world poll. Preliminary analytical findings from a multivariate model interfaced with derived clusters are presented.

Key Words: Trinary, Trinomial, Partitioning, Clustering, Subset regression, GINI Index, Big Data

1. International Socioeconomic Data Retrieved and Available on the Internet

The Program on Global Environmental Change at United Nations University, as described in their 2014 report, uses the Inclusive Wealth Concept consisting of components related to (a) manufactured goods, such as roads, buildings, machines, equipment, (b) education and health, and (c) natural resources related to ecosystems and atmosphere.

The Gallup-Healthways Global Well-Being Index derived for 135 countries was used in our study. The countries were ranked on Happiness consisting of 5 factors relating to physical health, financial security, and community support in social relationships..

We also used the following data available to the public:

(1) GINI index which measures income inequality, a lower value indicating more equality, higher values indicating higher inequality; i.e., most of the income going to only a small proportion of residents. This index has usually ranged between 24% and 49%, lowest for Scandinavian countries, and highest for African countries often, between 63% to 70%.

(2) Ethnic and linguistic fractionalization, indicating diversity versus uniformity in ethnic composition and languages used by the population.

(3) Membership in G-20, a dichotomous 0/1 variable.

(4) Having large cities, defined asa having At least one city among the ;largest 100 cities, and how many.

(5) Level of social support, indicated by a rank 1-10.

1.1 Formulating the Analysis Matrix

The data sources did not have a uniform number of countries; some countries were in one database but not in others. In order to create a matrix of m-dimensions and N-countries, we used only those the 62 countries which had data for all of the variables.

1.2 Analytic Applications: Regression/Cluster Merging

Iterative clustering allocated to each of the 62 countries into one of 3 clusters. Cluster I had 20 countries, Cluster II had 17 countries, Cluster III had 25 countries.

In order to further determine the composition of determinants within the countries in each cluster, a multivariate linear regression analysis was applied within each cluster, using Happiness as the dependent variable. The results are summarized in Table 1. Attribute definitions are described in the right; the left section shows how the six attributes fared in terms of their contribution (Yes) or (No) in terms of regression coefficients when Happiness is used as the dependent variable. The statistical significance cutoff level was .05, indicated as (Yes) or (No) , also showing whether increase (+) or decrease (-) in the input values was associated with overall Happiness. Thus, given an individual country, if we look up its cluster, it is possible to use Table 1 as a reference to provide guidance on which variables appeared to contribute.

For example, Income Inequality, as measured by the GINI Index, implies that it is positively correlated with national happiness for Cluster 1, negatively related to happiness for Cluster 2, and shows no significance either way for cluster 3.

Ethnic Diversity is an important detractor from happiness in Clusters 1 and 2, but not so much in Cluster 3. The latter countries may tend to be more ethnically homogeneous or more comfortably accommodated to their diversity.

Linguistic Diversity is a very different matter in Clusters 1 and 2. In Cluster 1 it seems to offer some positive enrichment or at least diversion, whereas in Cluster 2 it seems to be disruptive or at least disquieting. In Cluster 3 it has no significant effect.

G-20 Membership (or not)is interesting in two respects. First, there are clear and distinct differences between Clusters 1 and 2 on whether the variable contributes to or detracts from national happiness. Second, people within each nation have similar perceptions of the relative desirability of G-20 membership regardless of whether or not their particular country is a member.

Major City Presence or absence is simply not a happiness factor in any cluster.

Social Support is the one universally significant component of national happiness worldwide. This refers to ability to rely on others. They may be withregard to safety, health, shelter, food and water, or anything else that helps to avoid deprivation, isolation and helplessness. The support may come just from friends and family or it may be supplemented by the society at large.

The lower section of the Table 1 shows coefficients which may be used to build regression equations for each of the three clusters. The first line shows the intercept, followed by regression coefficients and standard errors (in parentheses) for each of the six independent variables. The statistical significance of the respective regression coefficients are also included.

Results for the total group, without regard to clustering, are also shown in the last column. As much opportunity as we have for speculation and inference when we evaluate the findings based upon clustering, we have the lack of this opportunity when clustering was not used. No statistical significance is observed except for the last variable related to social support. On the other hand, using the indices shown in Columns 2, 3, and 4 corresponding to Clusters 1, 2, and 3 we can use the intercept shown in row 2 of the table and the regression coefficients (+) or (-) for building a model or equation for each of the 62 countries, partitioning our analysis so that countries in Cluster 1 use the model based on Cluster 1 using coefficients in column 2, those in Cluster 3 use the coefficients in column 3, and those in Cluster 3 use coefficients shown in column 4.

Table 1: Attribute Definitions and Statistical Significance of Regression Coefficients for Independent Variables When Data Were Analyzed Using Cluster Partitioning

Significance and Direction of COEFFICIENTS FROM CLUSTER BASED REGRESSION

p<.05	CL1 N=20	CL2 N=17	CL3 N=25
GINI	YES (+)	NO	YES (-)
ETH	YES (-)	YES (-)	NO
LING	YES (+)	YES (-)	NO
G20	YES (+)	YES (-)	NO
+POP	NO	NO	NO
SOC SUP	YES (+)	YES (+)	YES (+)

ATTRIBUTE DEFINITIONS

- GINI: Income Inequality –high value indicates more unequal
- Ethnic Fractionalization
- Linguistic " " "
- Membership in G 20
- At least 1 City among World's Largest 100 Cities
- Rank for Social Support

REGRESSION/ CLUSTER MERGING

	CLUST (1)	CLUST (2)	CLUST (3)	TOTAL
Intercept	3.93 (.46)	7.73 (.17)	7.05(.48)	6.45 (.43)
GINI INDEX	.07(.01) p<.001**	.005 .(01) p=.35	-.04(.01) p=.01*	-.02 (.01) p=.87
ETHNIC	-2.92 (.44) p<.001)**	-1.55 (.24) p<.001**	-.10(.50) p=.06	-.45 (.48) p=.36
LINGUIST.	1.79 (.40) p<.001**	-1.51 (.25) p<.001**	.88 (.47) p=.07	-.43 (.45) p=.34
G-20 MEMB.	.96 (.33) p.= .012	-.49 (.13) p<.01*	-.01 (.27) p=.98	.49 (.29) p=.09
+ POP.	-.17 (.35) p=.63	-0.12 (.11) p=.31	-.03 (.24) p=.91	-.19 (.26) p=.45
RANK SOC.	.26 (.04) p< .001)**	.11 (.02) p<.001**	.19 (.03) p<.001**	.16 (.04) p <.001)**

1.3 Comparative Precision Evaluations through Residuals

In this section we refer to the increased precision in modeling through decreased error in residuals, difference between observed values and those calculated from regression equations, as we compare the model derived for the total group with those of trinary cluster-based subsets . Table 2 summarizes our results.

Table 2. Comparison of Coefficient of Variation, R2 and Residuals for the Total Group with those Obtained through Tripartite Cluster Partritions

Group	Mean (n)	s.d.	R2	>+/-10% residual	+/-3-9% residual	<+/-3% residual
TOTAL	6.29 (62)	0.78	0.35	24 : 39%	24 :39%	14: 22%
C1	6.73 (20)	0.73	0.90*	0	7	13: 65%
C2	6.56 (17)	0.67	0.97*	0	1	16:94%
C3	5.77 (25)	0.58	0.73	1	9	15:60%

As we observe in Column 4, the coefficient of variation, R-squared for the total group was .35, indicating that only one third of the total variation was explained, while through cluster-based partitioning R2 was .90 for Cluster 1, .97 for Cluster 2 and .73 for Cluster 3, indicating much higher precision or resolution in the models.

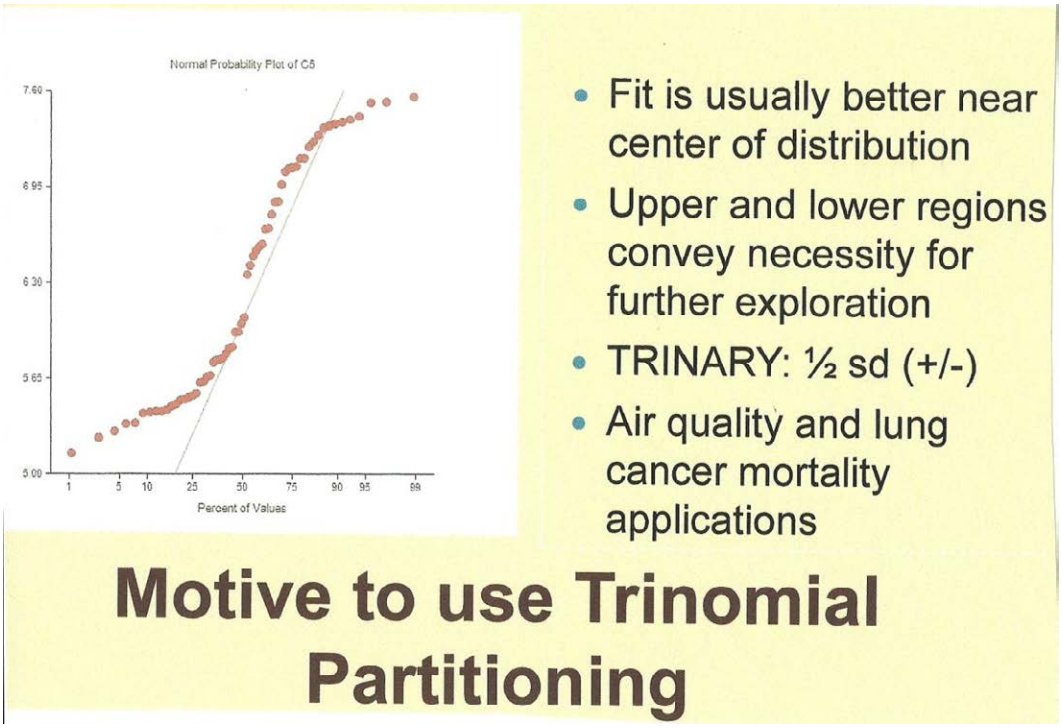
The last three columns of Table 2 show results show our results when the absolute value of residuals were divided into three groups: (a) +/1 10%,, (b) +/- (3-9%) and less than 3%. For the group as whole, when a single regression equation was fitted to the data, much less precision was observed. Of the 62 observations 24 (39%) had residuals exceeding 10%; for only 14 (22%) the residuals were less than 3%. On the other hand, the fit was more precise for the three clusters-- at least 60 % of the calculated values were within 3% of the observed values.

2. Recommendations Relative to Big Data Analytics

We are recommending a Trinary clustering approach interfaced with regression. Combining elements of cluster analysis with multivariate regression provides a more powerful tool not only for analysis but also for interpretation and tailoring models developed. Some prior observations based on analyses of research studies to justify our approach aree the following:

*2.*1 Observations Relating to Single Variable Plots

.



A cumulative probability plot, which is a straight line under Normality assumptions, often is not. The fit is usually better near tghe center of the distribution than in the upper and lower regions. In prior research, some including air quality and cancer mortality, we found that a Trinary (three category) approach delineating the center or "core" from non-core was beneficial,and that ½ of one standard deviation below and above the mean helped in this regard. Until now we dealt with bivariate analyses; in this paper we are presenting extensions to the multivariate domain.

2.2 Low Values Observed in Parametric and Non-Parametric Correlations.

Table 3 shows Pearson Product Moment and Spearman-Brown non-parametric correlations to pairs of variables used in the present study. In each bivariate entry the parametric Pearson Product Moment correlation coefficient is shown first, followed by the non-parametric Spearman-Brown correlation coefficient. Generally they are in agreement , but are quite low. Remembering that the square of the correlation coefficient is what explains the proportion of the attributable variance, most of the values do not explain even one haff of the total variability.

Table 3. Pairwise Pearson Product Moment and Spearman-Brown Correlations Applied to Variables in the Present Study

PEARSON / SPEARMAN
(in each entry)

	GINI	ETHN	LINGU	HAPPI	G 20	100 CIT	SOCIA
GINI	X	.31 /.37	.02/0.0	-.26/.30	0.0/.02	.26/.27	-.36/.45
ETHN		X	.53/.45	-.29/.32	-.10/.10	.11/.08	-.10/.19
LINGU			X	-.25/.23	-.13/.11	.04/.03	-.03/.03
HAPPI				X	.16/.14	-.06/.08	.48/.54
G 20					X	.68/.68	-.13/.13
100 CIT						X	-.20/.21
SOCIA							X

==========

In conclusion , use of clustering prior to multivariate regression in order to form subsets which will interface with multiple regression will enhance mathematical moodeling.