Reliability Analysis of the Commune Database (CDB) and its Income and Poverty Projection Model for Annually Poverty Monitoring at National and Sub-National Level

By: Ny Boret

MoI/NCDD/PST/M&E Unit

E-mail: nyboret@ncdd.gov.kh

H/P: (+855) 12 89 89 50

National Workshop-Phnom Penh Hotel, Sept 2009

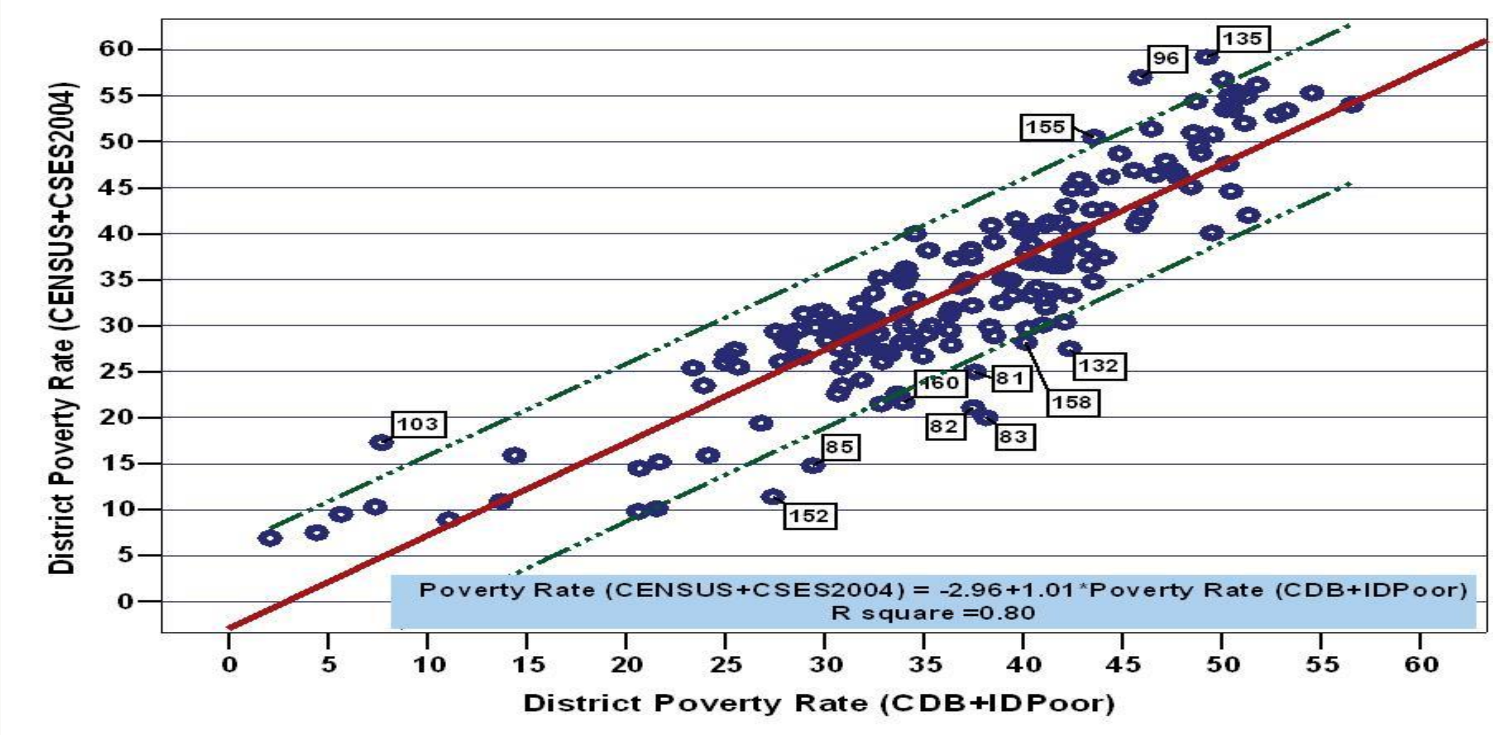
MoP/NCDD/UNDP

Objectives, Methods and Data used

- To measure the level of reliability of CDB and the Poverty projection model for long term poverty monitoring for National and Sub-National level
- There are to way to measure the level of Reliability of any database.
  - ***External consistencies check:*** *by simple cross checking the result from database with the result from other data sources, the disadvantage of this simple methods is that we do not know clearly where are the suspected object come from which data sources.* <u>*Other technique is using statistical modeling that can give the result more scientific and clearly picture (used by this study, CDB, IDPoor2007/08, CSES2003/04)*</u>
  - ***Internal consistencies check:*** *using one database alone with simple univariate statistic or* <u>*advance multivariate statistic techniques (used by this study CDB2006-2007)*</u>

Result: External consistencies check (Poverty projection model)

- Simple crossed check between predicted district poverty rate from modeling between (Census+CSES2004) and (CDB+IDPoor)

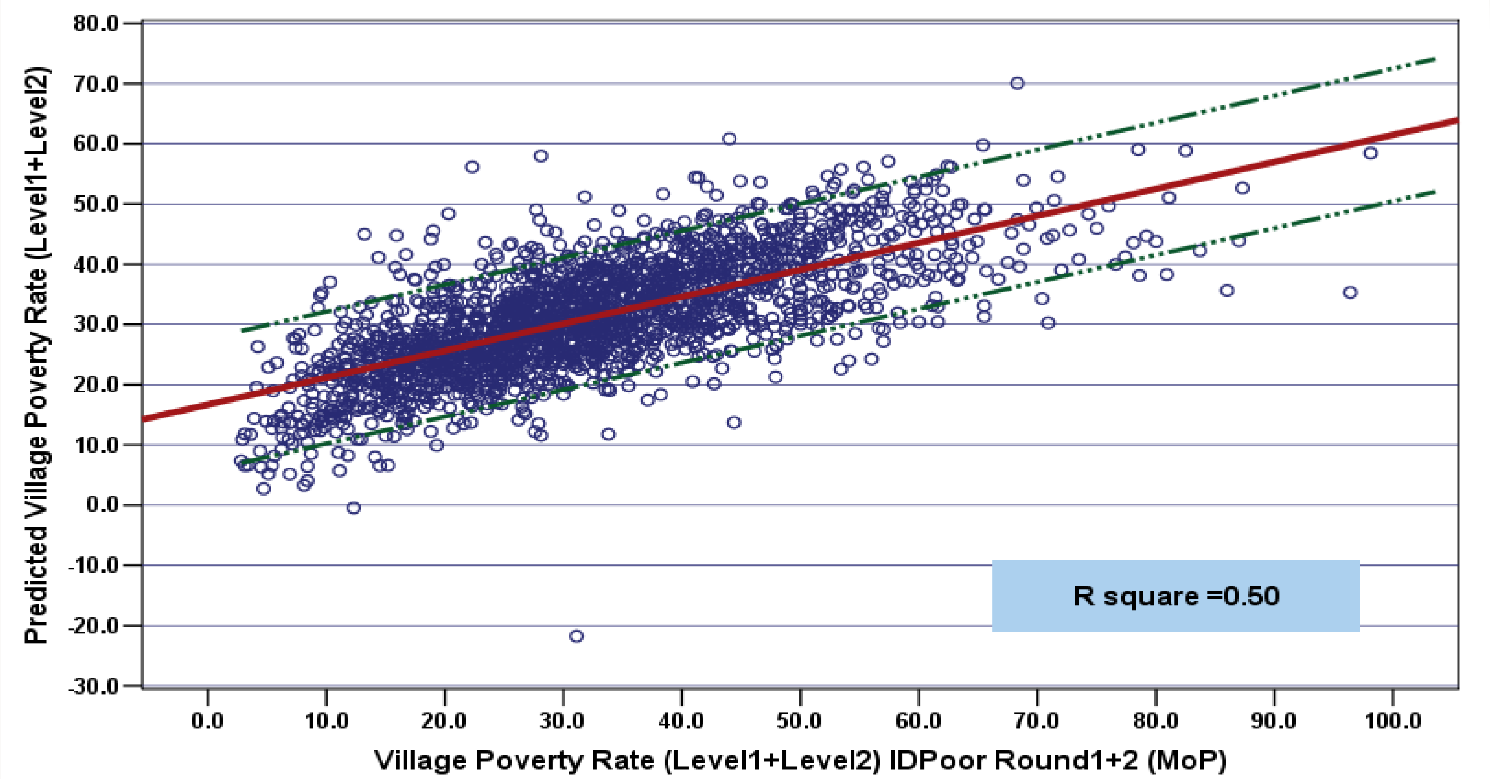


- As we can see only 10% of 185 predicted district poverty rate in 2004 come from (Census+CSES2004) and from (CDB+IDPoor) were not matched, these 15 districts are highly suspected of not accuracies in predicted poverty rate and it can be come from the prediction error from (Census+CESE2004) or the prediction error from (CDB+IDPoor). The disadvantage of this simple cross check is we do not know which is the source of error.

- ***Source:*** *District poverty rate (CENSUS+CSES2004) can be obtained from annex 1.2 page 61 of building local capacity in poverty mapping for Asian member counties. And district poverty rate (CDB+IDPoor) can be requested from the author.*

Result: External consistencies check (Poverty projection model) Con’t

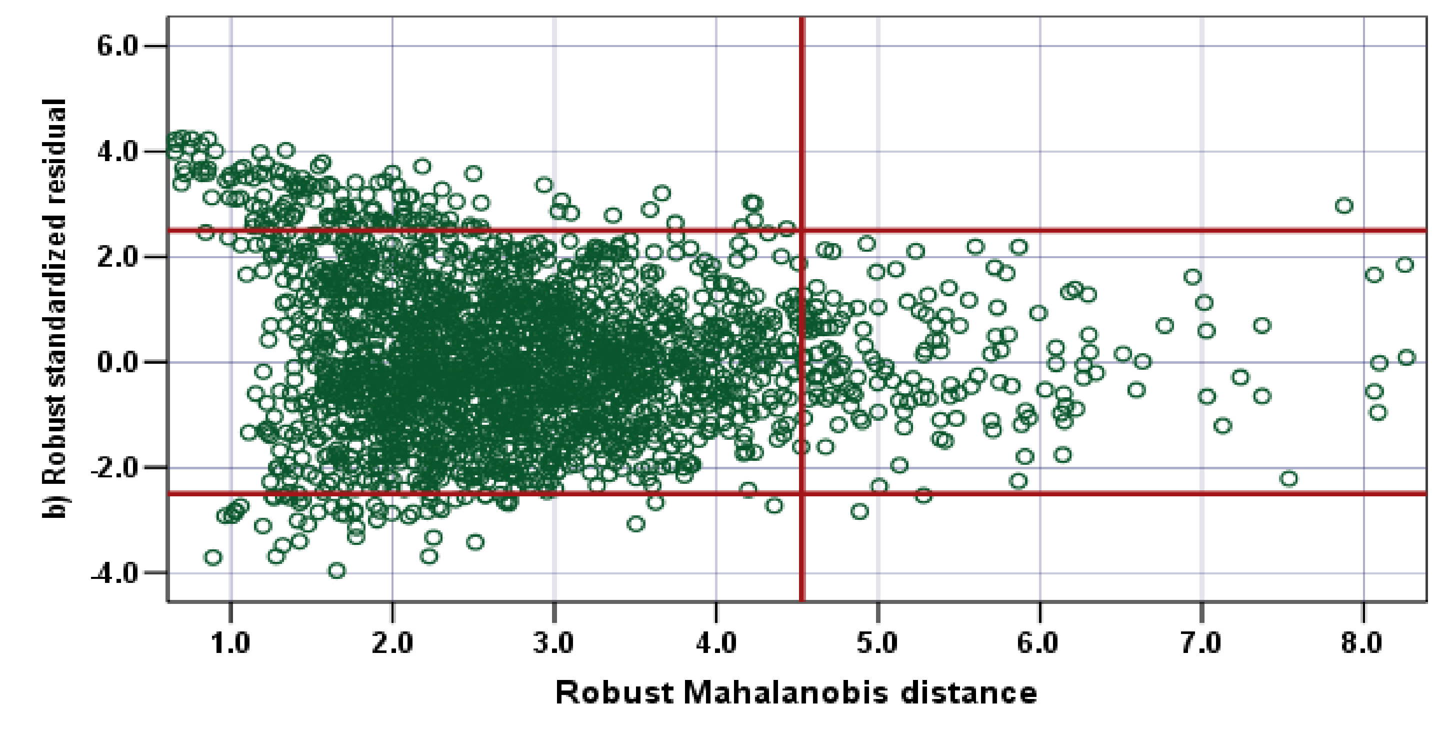
- *Scatter plot of predicted poverty rate and actual poverty rate between the CDB and IDPoor Round1+2*



- ***Note:*** **The sample village inside this scatter plot is (2106 sample villages) confirm that higher in poverty rate of IDPoor will be predicted to be higher in poverty rate by the poverty projection regression model using the CDB, about 50% of sample villages poverty are the same or little difference between the actual poverty rate of IDPoor and predicted poverty rate CDB, the remaining 38% are some difference and 12% are much differences between the actual poverty rate of IDPoor and Predicted poverty rate of CDB, this 12% we can call inaccuracies or suspected sample villages. The inaccuracies can be come from IDPoor or CDB, to find the suspected village come from IDPoor or CDB we use regression diagnostic technique to check.**
- ***Source:*** **author estimation based on robust multi-level regression modeling between IDPoor and CDB**

Result: External consistencies check (Poverty projection model) Con’t

- *Outlier and leverage observations (sample village)*



- ***Note:*** **The robust multi-level regression model of IDPoor and CDB is using Y (poverty rate) from IDPoor and independent variables (Xs) from CDB. So any sample village (out of 2106 scatter villages) that has absolute standardized residual larger than 2.5 at (Y axis) are call outlier or inaccuracies village and belong to the IDPoor (Totally 150 sample villages). Any sample village that has robust mahalanobis distance at (X axis) lager than the cut of value of 4.55 are call inaccuracies or leverage sample village and belong to the CDB (Totally 80 sample villages).**
- ***Source:*** **author estimation based on robust regression modeling between IDPoor and CDB**

Result: External consistencies check (statistical model) Con’t

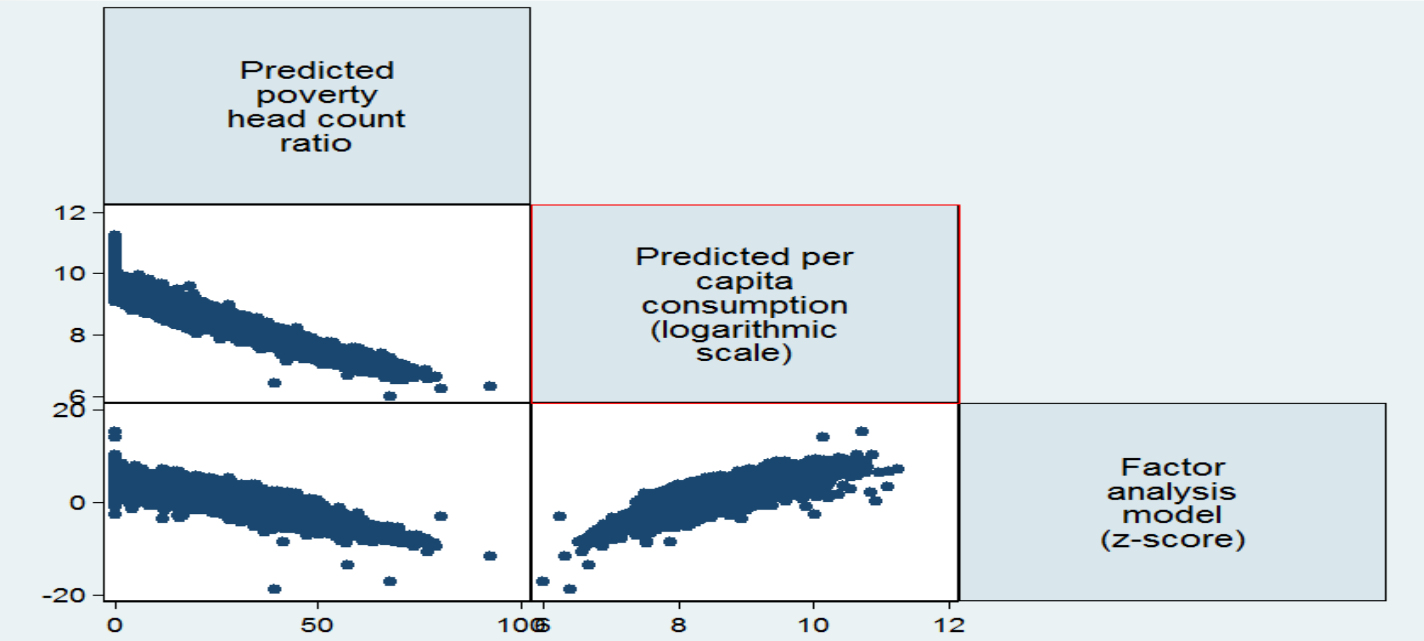
Technical note : Indicators and regression output at village level for poverty projection model (Model 1) using CDB2006-2007 and IDPoor round1+2 (MoP), Log expenditure projection model (Model 2) using (CDB2003 and CSES2004) and Factor analysis (z-score) model (Model 3) using (CDB2002-2008)

Independent variables	Model 1	Sig	Model 2	Sig	Model 3	Sig
	IDPoor2007/08		CSES2004		CDB2002-2008	
	Coefficent		Coefficent		Coefficent	
Intercept	24.19	0.00	8.62	0.00	1.68	0.00
Not latrine per family	10.45	0.00	-0.51	0.00	-0.81	0.00
TV per family	-5.17	0.00	0.41	0.00	0.27	0.00
Mountain/plateau region: 1=Yes, 0=Otherwise	3.50	0.00	-0.07	0.00	-0.69	0.00
Tonle Sap region: Reference	0.00	0.00	0.00	0.00	0.00	0.00
Plain region: 1=Yes, 0=Otherwise	-3.50	0.00	0.07	0.00	0.69	0.00
Coastal region: 1=Yes, 0=Otherwise	-7.00	0.00	0.13	0.00	1.37	0.00
Phnom Penh: 1=Yes, 0=Otherwise	-18.00	0.00	0.26	0.00	2.75	0.00
Urban area (exclude Phnom Penh): 1=Yes, 0=Otherwise	-5.00	0.00	0.14	0.00	2.79	0.00
Moto bike per family	-10.68	0.00	0.47	0.00	0.09	0.00
Household size	1.53	0.00	-0.10	0.00	-0.67	0.00
Concrete house per family	-6.37	0.40	0.27	0.00	2.97	0.00
Ratio of literate women18-64	-2.47	0.02	0.20	0.00	1.04	0.00
Ratio of men18-64 to all	-15.17	0.00	0.78	0.00	4.99	0.00
Thatch house per family	13.00	0.00	-0.49	0.00	-0.99	0.00
Bike cycle per family	-0.79	0.12	0.04	0.00	0.12	0.00
Ratio of house with electricity	-3.08	0.05	0.20	0.00	0.90	0.00
Ratio of family use TBA	26.57	0.00	-0.65	0.00	-5.64	0.00
Ratio of children 6-14 not go to school	2.20	0.05	-0.21	0.00	-1.46	0.00
Ratio water in home less than 150m	-1.54	0.05	0.10	0.00	0.78	0.00
N (number of observations village)	2106		900		92815	
Adjust R-SQUARE	0.50		0.70		0.54	

Note: Model 1 and Model 2 were adjusted non-normality, heteroscedasticity, outlier observation for both x , y spaces and intra cluster correlation (commune as random effect) by using robust multilevel (mixed effect) regression applying Bisquare weighted estimation

Result: External consistencies check (Poverty projection model) Con't

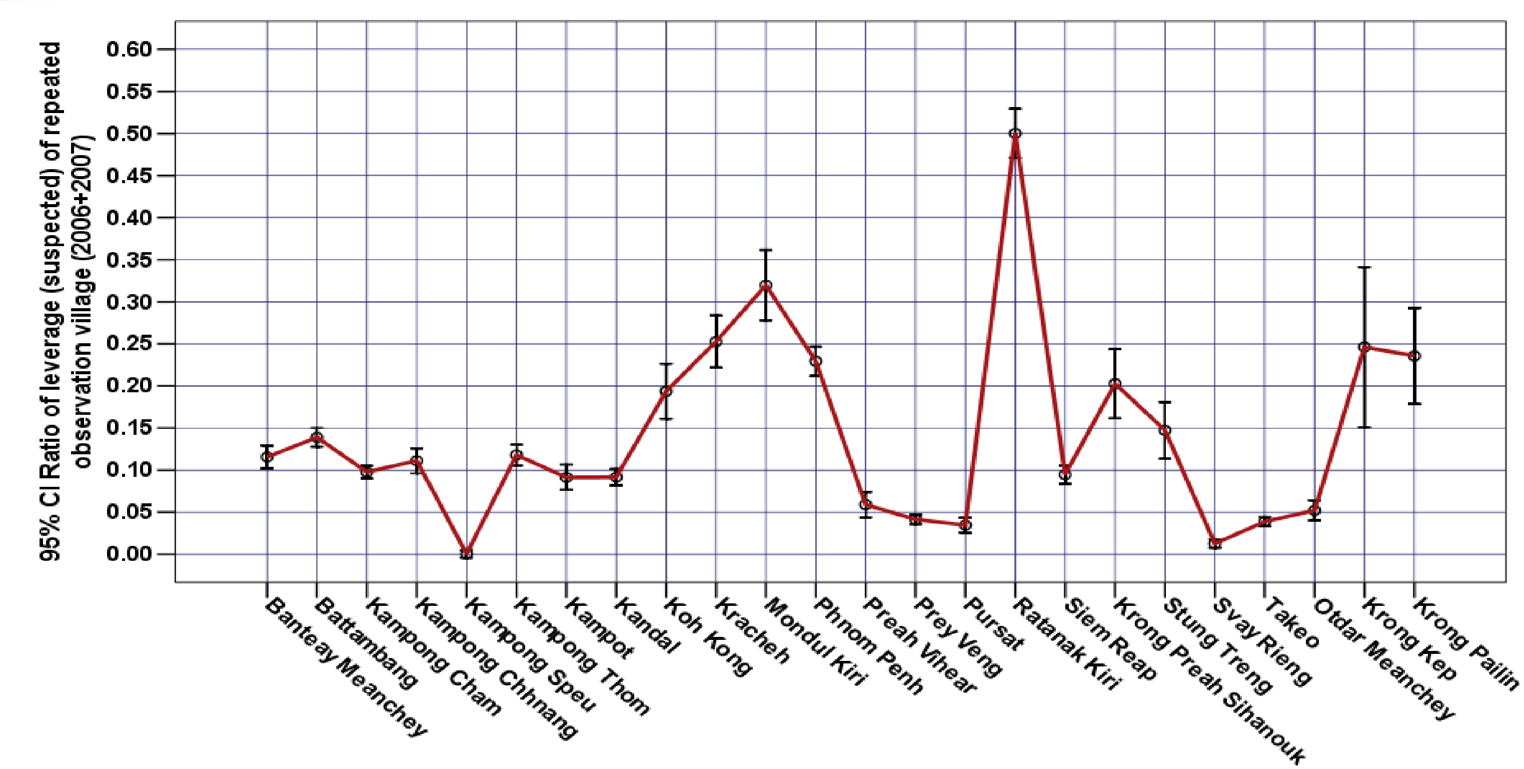
- *Accuracies of finding from the prediction model of regression and factor analysis model (scatter plot) for all villages 2002-2008*



- ***Note:*** **the accuracies of finding show us that, by using differences data source (IDPoor) 2007/08, (CSES) 2003/04 and CDB 2002-2008 come from difference cross sectional location and time, applying differences statistical techniques (Regression model between CDB and (IDPoor) 2007/08, between CDB and (CSES) 2003/04 and Factor analysis (PCA/FA) using CDB alone. This very large sample scatter plot of theses model are very significant correlated each other as in (scatter plot matrix = 92815 repeated observations village), this scatter plot confirm that the finding, the model and the proposed indicator are very accuracies and stable.**
- ***Source:*** **author estimation based on CDB 2002-2008**

Result: Internal consistencies check, multivariate statistical modeling (All of CDB variables as of ending 2007)

- *Ratio of suspected village or inaccuracies village (outlier or leverage observation village) by provinces, on average at national level 10% of repeat observation village are highly suspected to be as outlier*



- ***Note: Step 1:*** Using the Principle component and Factor analysis (PCA/FA) to reduce all indicators inside CDB2006-2007 in to a manageable indicators (now we call individual components score , each component score is the Z-score) that are well representative to all indicators and aggregate these Z-score to single index call total component score or total Z-score index
- ***Step 2:*** Using Regression analysis method let the total component score as the dependent variable and all individual component scores as the independent variable and applying regression diagnostic methods to find outlier (leverage observation) which observation are far away from the bulk of the data distribution that we will call suspected village.
- *Source:* author estimation based on CDB 2006-2007

Conclusions and recommendations

- CDB is a very good and cost effectiveness data sources for poverty monitoring for both at National and sub-nation and the only one data source for poverty monitoring and other CMDGs indicator at sub-national on yearly basis.
- Despite the average level of reliability is very high by external check with other database and also very high with the internal consistencies check is around (90% accuracies) , some province are still contained with high percentages of outlier village (suspected) village. I recommend more carefully recheck those villages and/or indicators during the data collection and data cleaning process before the data is compile into database. working with very large database like CDB appropriate statistical technique and computer software must be used for data cleaning and analyzing.

The End
Thanks for your attention