Towards an Understanding of Vulnerability in Rural Kenya

Luc Christiaensen (World Bank)
PADI Conference on Measuring, Understanding and Alleviating Household Vulnerability, Dar es Salaam, 2-3 February, 2004

1. What are we trying to achieve?

- ✓ Discuss and illustrate a methodology to understand and measure vulnerability
- ✓ Identify the determinants of vulnerability in rural Kenya
- ✓ Provide guidance on future data collection to study vulnerability as well as policies to reduce vulnerability in rural Kenya

How will we proceed?

2. Empirical strategy to measure vulnerability
   pseudo-panel and information on shocks

3. Rural households in Kenya
   risk factors, risk exposure and coping capacity

4. Empirical findings and policy simulations

5. Concluding remarks
   data collection and policy implications

2. Measuring vulnerability empirically

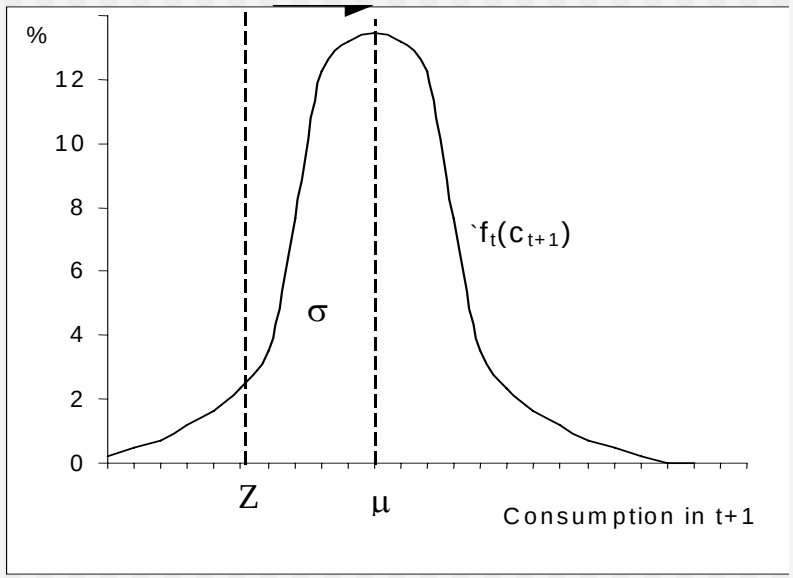
- ✓ Vulnerability as a measure of expected poverty
- ✓ Mathematical and graphical expression

$$V_{i,t,\gamma} = \int_{\underline{c}_{t+1}}^{z} (z - c_{t+1})^{\gamma} f(c_{i,t+1}) dc_{t+1}$$

If γ=0, V_0 = prob of shortfall

*If γ=1, V_1 =prob of shortfall * conditional expected gap*

*If γ=2, V_2=prob of shortfall * conditional expected gap squared*



Choices in measuring vulnerability

- ✓ Time horizon over which future shortfalls are assessed ➔ we take one period ahead.
- ✓ Indicator of well-being ➔ consumption
- ✓ Threshold for consumption ➔ poverty line (z)
- ✓ Estimation of probability distribution $f_t(c_{t+1})$
- ✓ Threshold for vulnerability such that persons whose expected poverty measure exceeds the threshold can be classified as vulnerable (e.g. 50%)

Key steps to estimate $V\gamma$ in practice

- ✓ Determination of a household's prob. distribution of fut. cons. $f_t(c_{t+1})$ is the major challenge
- ✓ If distribution lognormal ➔ knowledge of mean and variance suffices to know the distribution
- ✓ Given each household's ex ante mean and variance we can construct its ex ante distribution of future consumption
- ✓ Given each household's ex ante distribution we can estimate each household's probability of future shortfall $V\gamma$ given a certain poverty line

Key steps to estimate $V\gamma$ in practice (2)

- ✓ How to obtain the ex ante mean and variance?
  - ➢ We estimate the relationship between the ex ante mean and the ex ante variance of each household's future consumption and its current characteristics, those of its locality and the shocks it experienced during this period
  - ➢ Using these estimated coefficients together with household and location characteristics and the occurred shocks we predict the mean and variance of future consumption for each household
  - ➢ To do so, at least 2-period panel data needed (i.e. information on a household's future consumption and its current characteristics).

Consumption generating process (1)

- ✓ Households live in environments characterized by risks (e.g. droughts, sickness, price changes, violence/theft).
- ✓ Given risks, households allocate endowments to activities generating income. No perfect mapping from income into consumption b/c ex post consumption smoothing (asset depletion, borrowing, formal and informal insurance, migration, de-investment in human capital). The extent to which they can smooth their consumption = coping capacity.
- ✓ Given risks and in absence of sufficient coping capacity, households reduce risk exposure through ex ante income smoothing (income diversification, low risk-low return activities)

➔ risks, risk exposure and coping capacity jointly determine people's mean and variance of future consumption and thus their vulnerability

Consumption generating process (2)

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}\gamma_j + S_{ijt+1}\varphi_j' X_{ijt} + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j) \quad (1)$$

C_{ijt+1}	consumption of household i in area j at t+1
X_{ijt}	observable characteristics of household i and its environment j at t
S_{ijt+1}	observable locally covariant (weather shock) and idiosyncratic shocks (e.g. illness of income earner) experienced by household i in area j during t+1
θ_{ij}	time invariant unobservable household and environment characteristics
$\beta_j, \gamma_{j,}, \varphi_j, \alpha_j$	time invariant area specific coefficients to be estimated

Mean and variance of consumption

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}\gamma_j + S_{ijt+1}\varphi_j' X'_{ijt} + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$$
$$= X_{ijt}\beta_j + S_{ijt+1}(\gamma_j + \varphi_j' X'_{ijt}) + u_{ijt+1}$$

$$E(\ln C_{ijt+1} \mid X_{ijt}) = X_{ijt}\beta_j + \mu_S(\gamma_j + \varphi_j' X'_{ijt}) + E(\theta_{ij})$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_u$$
$$= \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt};\alpha_j)$$

- We implicitly assumed univariate shock, but if more than one shock (e.g. locally covariant weather shock and idiosyncratic illness shock) variance can be further decomposed in variance derived from covariant and idiosyncratic shocks

Mean and variance of consumption (2)

$$E(\ln C_{ijt+1} \mid X_{ijt}) = X_{ijt}\beta_j + \mu_S(\gamma_j + \varphi_j' X'_{ijt}) + E(\theta_{ij})$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt}; \alpha_j)$$

Data needed to obtain indiv estimates of mean and var of C_{ijt+1}

- info on household and locality characteristics $X_{ijt,}$
- historical information on shocks to derive the mean, variance and covariance of the shocks, $\mu_{S,}$ σ_S
- Estimates of $\beta_{j,}$ $\gamma_{j,,}$ $\varphi_{j,}$ $\alpha_{j,}$ σ_e
  - info on actual shocks experienced S_{ijt+1}
  - at least two period panel data with info on C_{t+1} and X_{ijt}
  - Assumptions on nature of unobserved θ_{ij} will determine $E(\theta_{ij})$, and σ_θ

Estimation methodology

Just and Pope (1978) or Amemiya (1977)

Step1: OLS regression of C on X and S

Step 2: use **squared estimated error term as estimate of variance of consumption** and regress it on X

Step 3: OLS regression of C on X and S corrected for heteroskedasticity to increase efficiency

Methodological issues - heteroskedasticity

- ✓ Household characteristics may influence expected cons. levels and volatility of cons. differently
- ✓ Usual specification ignores this possibility: $\ln C_{ijt+1} = X_{ijt}\beta_j + e_{ijt+1}$

  $E(\ln C_{ijt+1} | X_{ijt}) = X_{ijt}\beta_j$

  $V(\ln C_{ijt+1} | X_{ijt}) = \sigma_e$

  ➔ variance the same for all households

- ✓ Flexible heteroskedastic specification: $\ln C_{ijt+1} = X_{ijt}\beta_j + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$

  $E(\ln C_{ijt+1} | X_{ijt}) = X_{ijt}\beta_j$

  $V(\ln C_{ijt+1} | X_{ijt}) = \sigma_e h(X_{ijt};\alpha_j)$

  ➔ variance differs for each household depending on X_{ijt}

  ➔ X_{ijt} may affect mean and variance of consumption differently

  e.g. assets may increase mean and decrease variance

Methodological issues (2) – data on shocks

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}\gamma_j + S_{ijt+1}\varphi_j' X'_{ijt} + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt};\alpha_j)$$

- ✓ W/o data on locality covariant shocks, it is difficult to obtain a good estimate of the variance of consumption.
- ✓ Cross-sectional hh data combined with historical data on locality covariate shocks can be used to estimate variance of consumption.
- ✓ Spatially disaggregated data on the incidence of idiosyncratic shocks in different areas can be used similarly
- ✓ Interactions of household characteristics with observable shocks can be used to assess whether ability to cope with shocks varies systematically across households

Methodological issues (3) cross-sectional, pseudo-panel or panel data

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}(\gamma_j + \varphi_j' X'_{ijt}) + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$$

$$E(\ln C_{ijt+1} \mid X_{ijt}) = X_{ijt}\beta_j + \mu_S(\gamma_j + \varphi_j' X'_{ijt}) + E(\theta_{ij})$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt};\alpha_j)$$

✓ **Cross-sectional (CS) data**

- assumes X_{ijt} relatively constant over time ➔ C_t on X_{ij} and S_{ijt}
- unobserved heterogeneity (θ_{ij}) may bias estimates of β_j,γ_j and φ_j and thus less appropriate for exploring causes of vulnerability.
- to the extent its effect captured by β_j,γ_j and φ_j it will not bias the predicted mean of consumption
- unobserved heterogeneity in the cross-section confounded with intertemporal variation leading to overestimates of variance

➔ Overall, in the absence of panel data and when **combined with information on shocks** and sufficient variation in covariant shocks, CS can be quite **useful in providing a vulnerability profile**, though not so good to explore the causes of vulnerability

Cross-sectional, pseudo-panel or panel data

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}(\gamma_j + \varphi_j' X'_{ijt}) + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$$

$$E(\ln C_{ijt+1} \mid X_{ijt}) = X_{ijt}\beta_j + \mu_S(\gamma_j + \varphi_j' X'_{ijt}) + E(\theta_{ij})$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt};\alpha_j)$$

✓ **Panel data**

- equivalent to differencing and thus permit to control for unobserved heterogeneity

➔ unbiased estimates of coefficients, the mean and variance

➔ given several observations on shocks on same households you obtain more accurate estimation of γ_j and φ_j

- problem of measurement error (ME) in C_{ijt}, more severe than in cross-section➔ overestimates of variance

- panel data typically not widely available and limited in their coverage

- availability of panel data does not really compensate for lack of information on shocks, unless the panel is long

➔ Panel data preferred and **necessary for causal analysis of vulnerability**; its collection should be encouraged though issue of ME

Cross-sectional, pseudo-panel or panel data

$$\ln C_{ijt+1} = X_{ijt}\beta_j + S_{ijt+1}(\gamma_j + \varphi_j' X'_{ijt}) + \theta_{ij} + e_{ijt+1} h^{1/2}(X_{ijt};\alpha_j)$$

$$E(\ln C_{ijt+1} \mid X_{ijt}) = X_{ijt}\beta_j + \mu_S(\gamma_j + \varphi_j' X'_{ijt}) + E(\theta_{ij})$$

$$V(\ln C_{ijt+1} \mid X_{ijt}) = \sigma_S(\gamma_j + \varphi_j' X_{ijt})^2 + \sigma_\theta + \sigma_e h(X_{ijt};\alpha_j)$$

- ✓ **Pseudo-panel data**

- pseudo panels are panels of groups (vs individuals) defined by time invariant characteristics (age, location, education of head,...)

-estimates of coefficients are obtained by examining the relation between group means as opposed to the relation between consumption and characteristics of individual households

- pseudo panel data can be constructed from repeated cross-sections (e.g. age cohorts) or through repeated random sampling of households from the same clusters

Pseudo-panels data (2)

Given a grouping of households in cohorts c, eq (1) becomes:

$$\ln C^{*}_{cjt+1} = X^{*}_{cjt}\beta_j + S^{*}_{cjt+1}(\gamma_j + \varphi'_j X^{*'}_{cijt}) + \theta^{*}_{cj} + e^{*}_{cjt+1} h^{1/2}(X^{*}_{cjt};\alpha_j)$$

- * indicates cohort population means which in practice will be replaced by sample means
- as a result θ^{*}_{j} not constant over time given different individuals each period, but the larger the sample size, the closer it approaches the population mean ➔ cohort fixed effects (differencing) permits you to control for unobserved heterogeneity and obtain unbiased estimates of coefficients and the mean and variance
- ME problem is less severe because by taking the cohort means, measurement errors are largely averaged out if ME random (not correlated to cohort characteristics) and zero on average
- Group vs individual estimators?
  - ➔ efficiency increases the more homogeneous the groups
  - ➔ need for a sufficient number of groups, but often several repeated cross-sections present
  - ➔ correction for taking sample vs population means

How will we proceed?

2. Empirical strategy to measure vulnerability
   pseudo-panel and information on shocks

3. Rural households in Kenya
   risk factors, risk exposure and coping capacity

4. Empirical findings and policy simulations

5. Concluding remarks
   data collection and policy implications

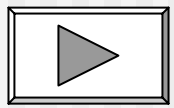
3. Household vulnerability in rural Kenya

- ✓ Vulnerability profile based on national 1994 and 1997 WMS
- ✓ 1994 as current and 1997 as future period
- ✓ Vulnerability closely linked to people's livelihood systems. We distinguish:
  - ➢ Urban communities
  - ➢ Rural pastoral communities
  - ➢ Rural non-pastoral communities ➔ our focus
    - Arid and semi-arid
    - Humid and sub humid zones

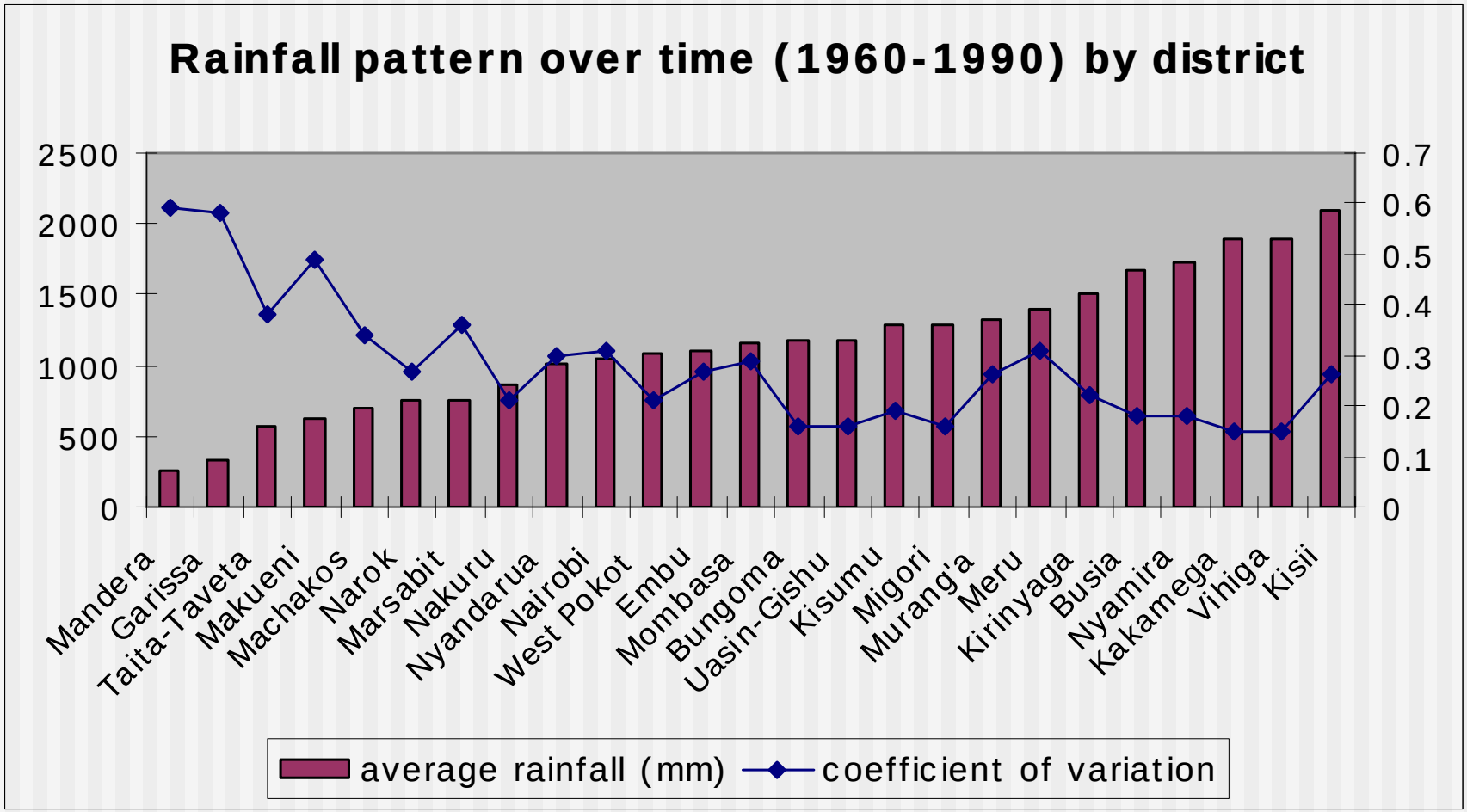
Empirical specification

- ✓ Hybrid between cross section and pseudo-panel
  - ➢ Repeated survey of random sample of households from same communities (clusters) in 1994 and 1997
  - ➢ Only two observations → thus not possible to apply panel estimation techniques, unless one is willing to assume X time invariant
- ✓ Use of group means reduces bias in estimating the variance due to measurement error
- ✓ Given that individuals in both years in each group not all the same, bias due to unobserved heterogeneity also less problematic

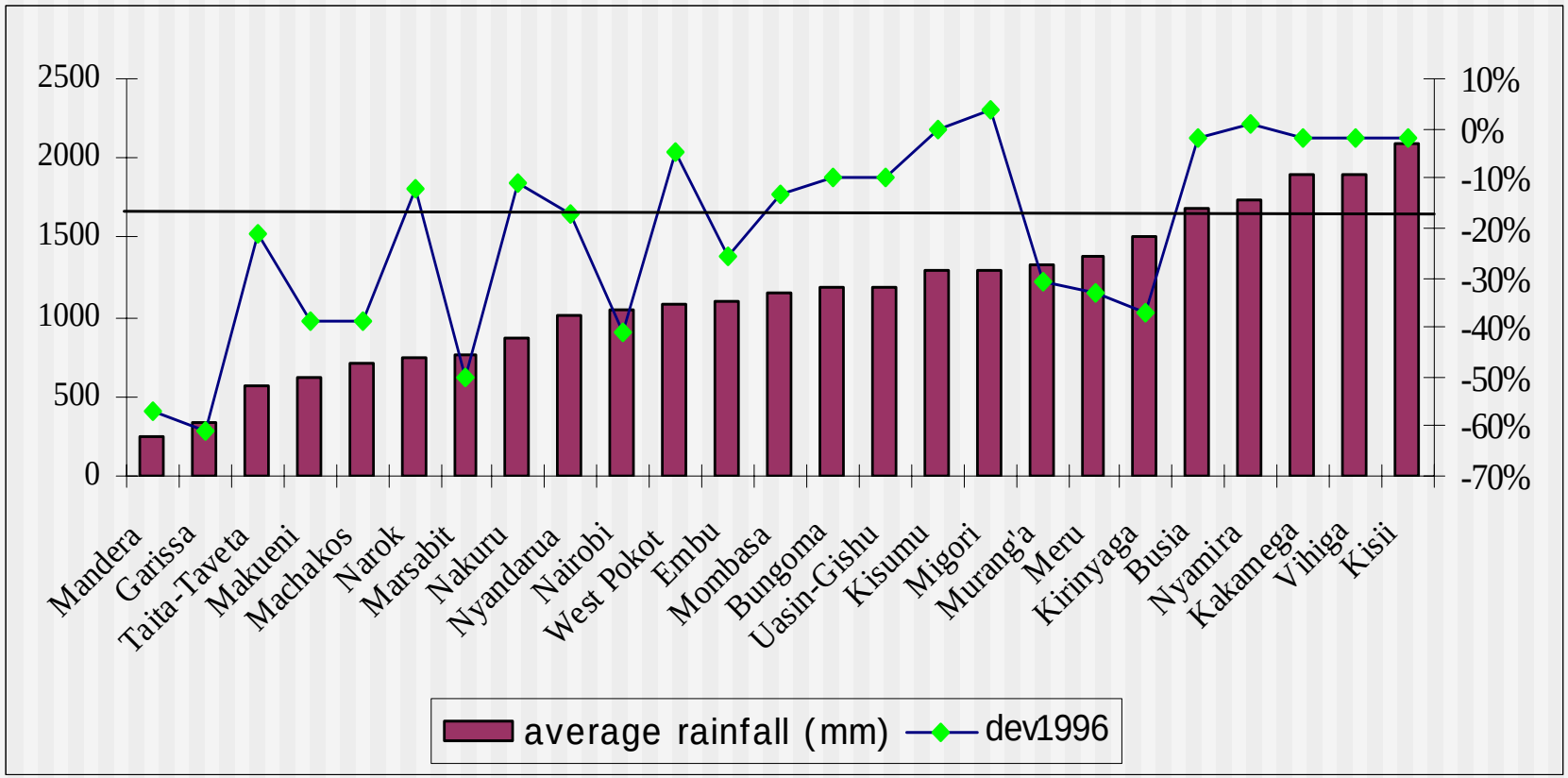
Choice of covariates

- ✓ Dependent variable: real 1997 expenditure/adult equiv.
- ✓ Risk factors:
  - ➢ 1996 rainfall shock; malaria incidence during last 2 wks in 1994 
- ✓ Risk exposure (1994)
  - ➢ Landholdings, fertilizer use, proxies for income diversification
- ✓ Coping capacity (1994)
  - ➢ Household size, dependency ratio,gender household headship, adult literacy rate
  - ➢ Number of sheep/goats and cattle
  - ➢ Use of electricity
  - ➢ Time to food market 

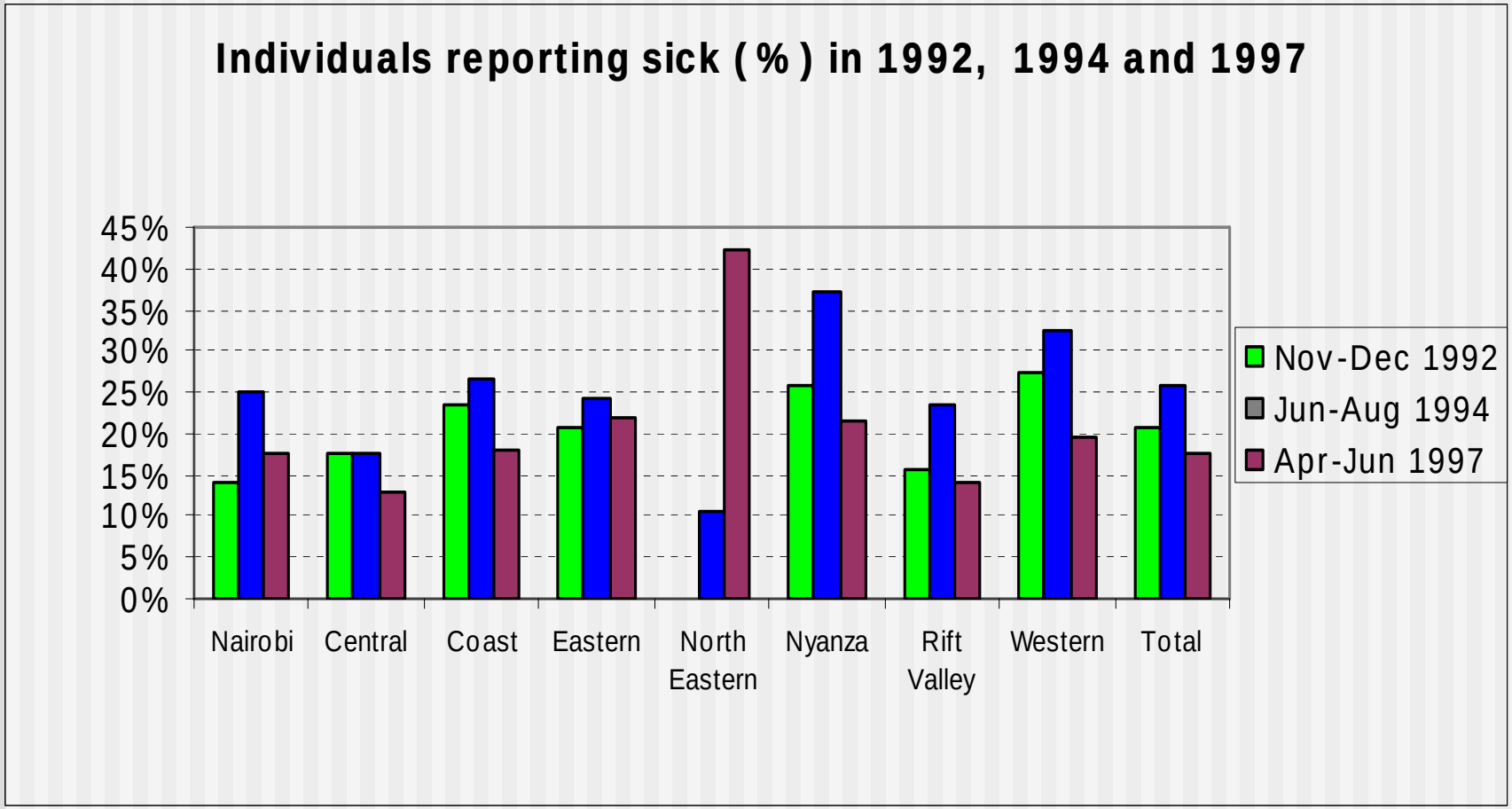
Major risk factors: Rainfall



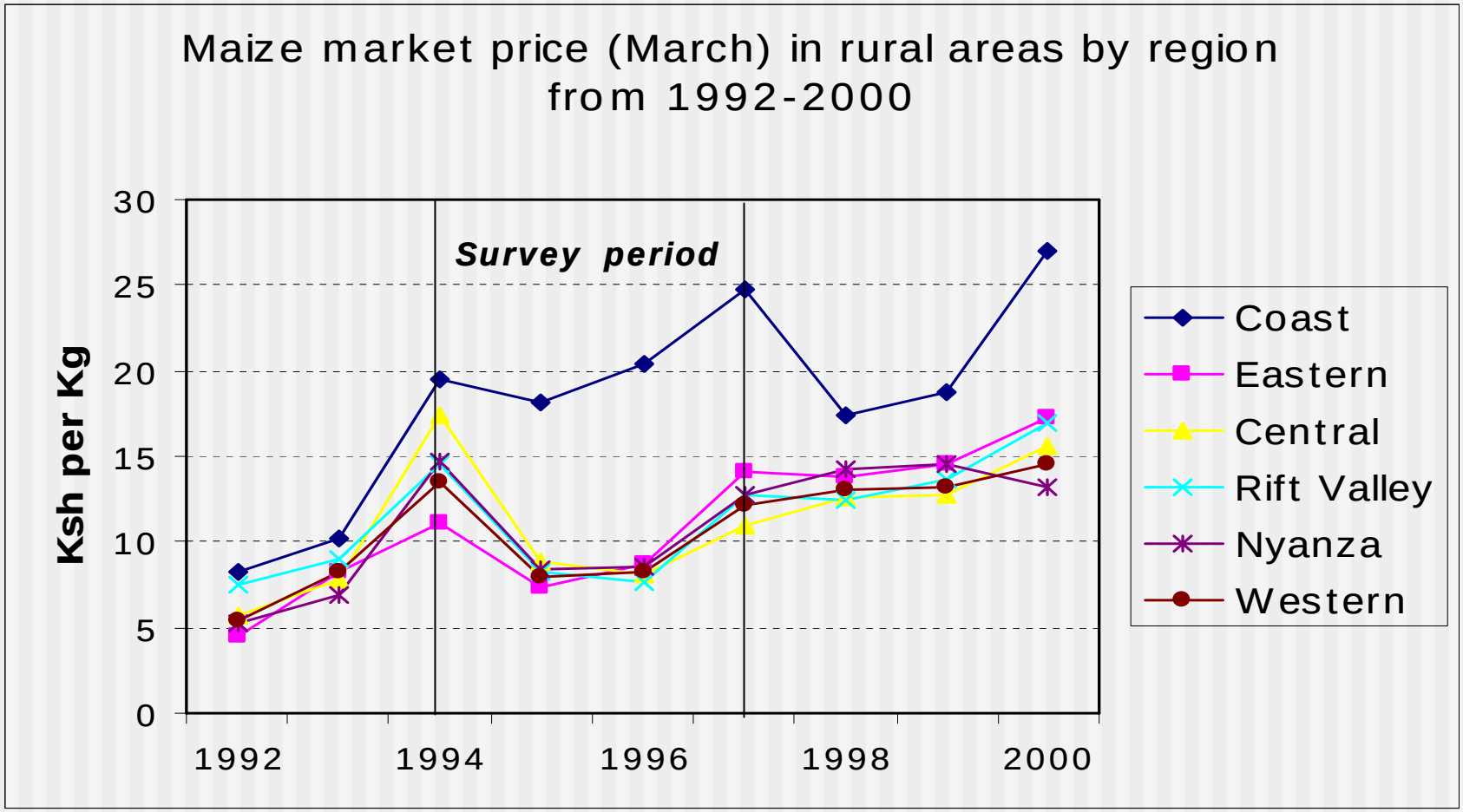
Mean rainfall (1960-90) and deviation from mean (%) in 1996



Sickness incidence



Maize Market Price



How will we proceed?

2. Empirical strategy to measure vulnerability
   pseudo-panel and information on shocks

3. Rural households in Kenya
   risk factors, risk exposure and coping capacity

4. Empirical findings and policy simulations

5. Concluding remarks
   data collection and policy implications

4. Vulnerability profile by location

Vulnerability profile of non-pastoral rural communities in 1994

Province (# observations)	Prob of shortfall (V0)	Proportion V0>0.5
Central (143)	0.23	0.03
Coast (71)	0.40	0.30
Eastern (121)	0.38	0.12
Nyanza (164)	0.45	0.26
Rift Valley (21	0.33	0.09
Western (94)	0.41	0.17
National	0.36	0.15
# observations	808.00	808

Determinants of vulnerability

- ✓ Risk factors
  - In *non-arid* areas households/communities with a larger incidence of malaria, face on average a larger probability of shortfall (lower mean and larger variance); effect not significant in *arid* areas
  - In arid areas, rainfall shocks negatively affect probability of shortfall, though effect not significant in non-arid areas

Determinants of vulnerability (2)

- ✓ Risk exposure
  - Access to non-farm employment greatly enhances mean of future consumption and reduces the variance in both arid and non-arid areas
  - Further investigation of non-farm employment in non-arid areas shows:
    - Especially skilled private sector workers less vulnerable (higher mean and lower variance)
    - Unskilled private sector workers also higher mean, though also higher variance (less secure income source)
  - Pensions are variance reducing, though no effect on mean in non-arid zones

Determinants of vulnerability (3)

- ✓ Coping capacity
  - Possession of cattle increases the mean of households' consumption, especially in arid areas, though no effect on the variance;
  - Possession of sheep/goat tends to reduces the variance (more liquid, less correlated to shocks) especially in arid zones (protection against idiosyncratic shocks).
  - Electricity access improves average consumption
  - Market accessibility decreases vulnerability through mean effect – effect four times larger in arid compared to non-arid zones (higher food prices, lower cash crop prices, less diversification out of riskier low yielding food crops into high return cash crops (Omamo, 1998)
  - Literacy increases average consumption and decreases variance in non-arid areas

Vulnerability reducing potential of interventions

%change after intervention	Probability of shortfall (V0)	
	Non-arid	Arid & semi-arid
all communities <=10% malaria/fever incidence past 2wks		
all non-arid communities>=5% of adult members as skilled worker in pvt sector and >=10% as unskilled worker; arid&semi-arid communities >=25% of inc from non-agriculture activities		
All communities >=75% adult members literate		
all communities within 30 minutes from market		

vulnerability can be substantially reduced among rural non-pastoral communities ➔ role of public workfare programs

6. Concluding remarks

- ✓ Data collection
  - ➢ Critical importance of current and historical information on major shocks (rainfall, illness, prices)
  - ➢ Desirability for further panel data collection, though not always easy at national level; in the absence of household panels, panels of communities may suffice

- ✓ Policies.
  - ➢ Rural non-pastoral communities relatively vulnerable: in 1994 15 % ≥ 50 % chance to fall below the poverty line in the future. Substantial difference by province ➔ targeting
  - ➢ Promising interventions to reduce vulnerability are :
    - Targeted interventions to reduce malaria incidence
    - Improved access to food markets
    - Enhancement of adult literacy ratio
    - Promotion of off-farm employment opportunities

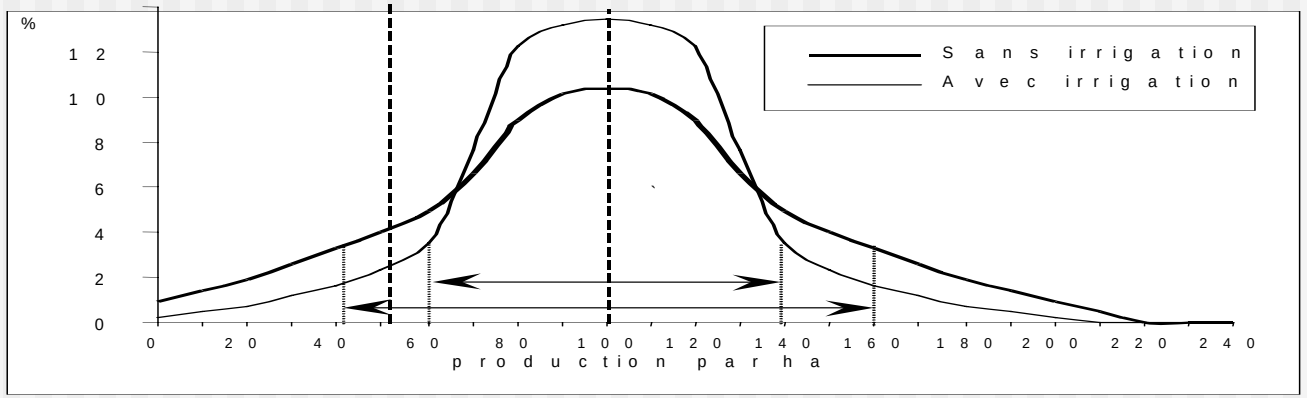
Timeliness of empirical results

- ✓ Timeliness of vulnerability profile depends on temporal stability of:
  - Estimated returns/coefficients to household and locality endowments
  - Rainfall pattern: vulnerability profile reflects 1996 rainfall pattern. Expected value of vulnerability measures based on probability of rainfall shocks indicates that basic findings of vulnerability profile remain, though we find a lower average probability of consumption shortfall in the arid zones (29 instead of 34 %).
  - Household and locality endowments ➔ periodic update

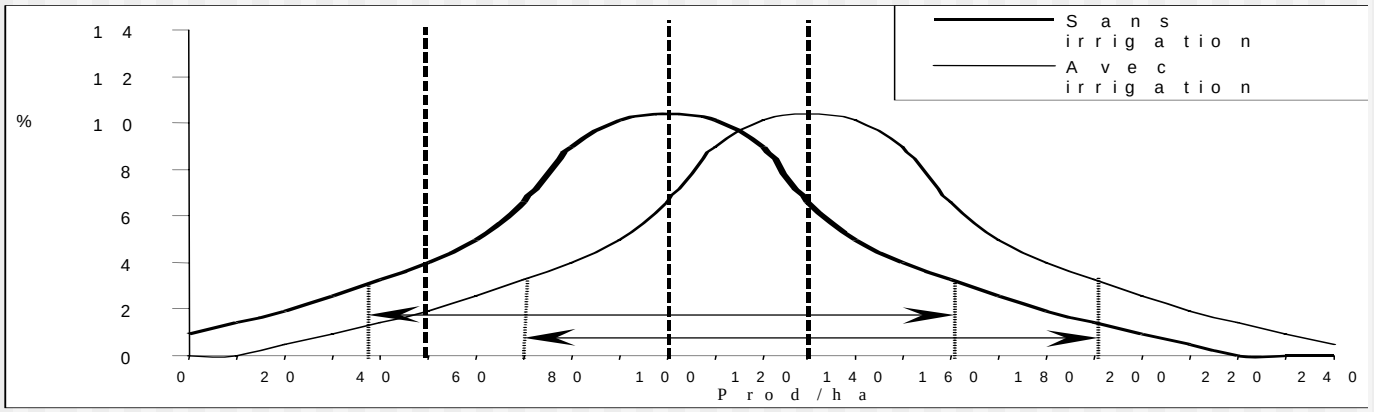
➡ 1994 profile good basis for characterizing and locating vulnerable communities

L'irrigation influence rentabilité et variabilité

Variabilité



Productivité



5. Empirical findings, cont. 2: Vulnerability profile by demography and occupation

Vulnerability profile of non-pastoral rural communities in 1994

Probability in 1994 of becoming poor in the future (1997) (=V_0)	Quintiles (Q1=low ; Q5=high)	Average prob of cons. Shortfall	Ratio of Q5/Q1
Demographic (average per community)			
proportion children (0-7 yrs old)/household	Q1	0.39	0.62
	Q5	0.24	
proportion children (8-15 yrs old)/household	Q1	0.24	1.79
	Q5	0.43	
proportion active female adults/ household	Q1	0.34	1.18
	Q5	0.40	
Proportion female headed households	Q1	0.33	1.27
	Q5	0.42	
Occupational (average per community)			
% adults/household with commercial farming as main occupation	Q1	0.37	0.86
	Q5	0.32	
% adults/household with subsistence farming as main occupation	Q1	0.23	2.04
	Q5	0.47	

5. Empirical findings: robustness of results, cont. 3

- ✓ Robustness of results:
  - Correlation between 1994 vulnerability measures V_0, V_1 and V_2 and actual average consumption in 1997 is –0.47, -0.44, --0.42 respectively
  - Average expected shortfall (= 1350 Ksh) is close to actual average consumption gap (=1255 Ksh)
  - Two thirds of communities correctly classified either as non-poor or as poor by vulnerability measure (probability cut-off of 50 %)

➔ Vulnerability measures emerge as solid predictors of future poverty