Análise do Grau de Precisão para os Resultados do Censo Agro-Pecuário 1999-2000 de Moçambique, e Revisão dos Planos de Amostragem para o Trabalho de Inquérito Agrícola 2002

David J. Megill
Consultor de Amostragem

Março de 2002

1. Introdução

O Instituto Nacional de Estatística (INE), em estreita parceria com o Ministério de Agricultura e Desenvolvimento Rural (MADER), realizou o último Censo Agro-Pecuário (CAP) entre Outubro de 2000 e Março de 2001. O CAP é a fonte mais importante de indicadores agro-pecuários a nível de distritos administrativos, e também vai servir como base de amostragem para inquéritos agro-pecuários anuais, como o Trabalho de Inquérito Agrícola (TIA). Dada a limitação de recursos, a enumeração do CAP foi feita numa amostra de áreas representativas dentro de cada distrito do país. Dado que os resultados do CAP são sujeitos a erros amostrais, é importante obter medidas do nível de precisão. Os termos de referência do consultor incluem a análise dos erros amostrais para os resultados mais importantes do CAP.

O CAP vai servir de base para a selecção da amostra para o Trabalho de Inquérito Agrícola (TIA) 2002, planificado para começar em Junho de 2002. O TIA é um inquérito agro-pecuário compreensivo para medir o nível de produção de culturas e animais, como também características das explorações. Outro objectivo é de obter medidas socio-económicas dos agregados familiares rurais, para avaliar os programas de desenvolvimento do sector. Os termos de referência do consultor também incluem a revisão dos planos preliminares de amostragem para o TIA 2002, para avaliar a eficiência do desenho e estimar o nível de precisão aproximado que se pode esperar para os resultados.

No ámbito de assistência técnica prevista para a Direcção de Economia (MADER), sob direcção do Director Carlos Mucavel, esta consultoria foi financiada pelo USAID (*United States Agency for International Development*). As conclusões deste estudo foram elaboradas baseadas nas discussões sobre a metodologia do CAP e os planos preliminares para o TIA com o pessoal técnico e administrativo de INE, MADER e MSU. O consultor quer agradecer especialmente a colaboração de Domingos Diogo, Director Adjunto do CAP, Christopher Hill, *Senior Census Advisor* da FAO (*Food and Agriculture Organization*) das Nações Unidas, e os Economistas Agrários de Michigan State University: Jan W. Low, Duncan H. Boughton, David Tschirley e Higino de Marrule.

2. Desenho da Amostra para o CAP

A amostra do CAP foi realizada pelo INE com a assistência técnica de Chris Hill. A base de amostragem para o CAP foi derivada dos dados e cartografia do Recenseamento Geral da População e Habitação (RGPH) de 1997. Dada a importância de alguns indicadores agro-pecuários a nível distrital, a base de amostragem for estratificada por distrito. Moçambique está dividido em 146 distritos administrativos. Originalmente todos os 138 distritos com características rurais foram incluídos no CAP, mas dois distritos não foram inquiridos por causa das inundações. Então o número final de distritos cubertos pelo CAP foi 136.

No caso das grandes explorações, um marco especial foi elaborado baseado em listas administrativas, com um total de 429 explorações. Dada a maior contribuição por unidade das grandes explorações nos totais de área e produção de culturas e produção agro-pecuária, todas foram incluídas no CAP com uma probabilidade de 1 (isto é, são auto-representadas na amostra).

A amostra de pequenas explorações e médias para o CAP foi seleccionada em duas etapas de amostragem. Para propósitos de seleccionar as unidade primárias de amostragem (UPAs), usou-se as áreas de enumeração (AEs) definidas para o RGPH 1997, que contêm entre 120 e 150 agregados na área urbana, e entre 80 e 100 agregados na área rural. Na primeira etapa de amostragem, uma amostra de AEs foi seleccionada dentro de cada distrito sistemáticamente com probabilidade proporcional ao tamanho, onde a medida de tamanho foi o número de agregados baseado nos dados do RGPH 1997. Dado que em geral foi difícil identificar os limites dos AEs com a cartografia censal, foi decido de definir a UPA como a unidade administrativa mais pequena na qual cai a AE seleccionada; neste caso, a UPA pode ser um quarteirão, um bairro ou uma aldeia. Outra razão de usar unidades administrativas para UPAs é que o chefe local podia assistir com a elaboração da lista de agregados familiares dentro da UPA. O número de UPAs seleccionadas por distrito varia entre 8 e 40, dependendo do tamanho do distrito (em referência ao número de agregados familiares). Desta maneira um total de 2,808 UPAs foram seleccionadas.

Dado que as UPAs foram seleccionadas indirectamente através das AEs, é importante considerar as probabilidades de selecção da primeira etapa. Neste caso a probabilidade de uma UPA em particular ser seleccionada seria igual à suma das probabilidades de todas as AEs que são incluidos nesta UPA. Dado que as AEs foram seleccionadas proporcionalmente ao tamanho, a medida de tamanho da UPA seria a suma das medidas das AEs correspondentes.

Dentro de cada UPA seleccionada desta maneira, o chefe local assistiu na elaboração de uma lista de agregados familiares para a selecção de explorações correspondentes. Foram identificadas as medias explorações dentro da UPA, com 10 ou mais hectares; todas estas explorações foram incluídas na amostra na segunda etapa.

No caso das pequenas explorações (com menos de 10 hectares), seleccionou-se 8 agregados familiares dentro de cada UPA. Como se explica na seguinte secção, isto resulta numa amostra auto-ponderada (com factores de expansão iguais) dentro de cada distrito para as pequenas

explorações.

3. Factores de Expansão para a Amostra do CAP

Dado que as probabilidades de selecção das explorações na amostra do CAP variam por distrito e por tamanho de exploração, foi necessário calcular factores de expansão correspondentes. O factor de expansão básico é simplesmente o inverso da probabilidade final de selecção, com um componente de cada etapa de amostragem. No caso das pequenas explorações, a probabilidade de selecção se pode expressar como o seguinte:

$$p_{hi(p)} \text{ ' } \frac{n_h \times M_{hi}}{M_h} \times \frac{8}{M'_{hi}} \text{ ,}$$

onde:

$p_{hi(p)}$ = probabilidade de selecção de pequenas explorações na i-éssima UPA amostral no estrato (distrito) h

n_h = número de UPAs amostrais seleccionadas no distrito h

M_{hi} = número total de agregados familiares no marco amostral do RGPH 1997 para a i-éssima UPA amostral no distrito h (a soma do tamanho dos AEs dentro da UPA)

M_h = número total de agregados no marco amostral do RGPH 1997 para o distrito h (a soma das M_{hi} dentro do distrito)

M'_{hi} = número total de agregados familiares com pequenas explorações identificadas na listagem para a i-éssima UPA amostral no distrito h

No caso do CAP a medida de tamanho M_{hi} não é conhecida antes da selecção das UPAs. Como se explicou na secção anterior, as UPAs foram seleccionadas indirectamente pelas AEs. Neste caso, se pode dizer que M_{hi} é a soma do número de agregados familiares em todas as AEs que se encontram dentro da UPA. Também se pode ver nesta expressão para a probabilidade que no caso em que $M'_{hi} = M_{hi}$ (isto é, o número de agregados listados é igual ao número correspondente do RGPH 1997), a probabilidade é constante para o distrito. Neste caso, o numerador da probabilidade seria o número de agregados familiares com pequenas explorações seleccionadas no distrito, e o denominador seria igual ao número total to agregados familiares no distrito. Por esta razão os factores de expansão usados para as pequenas explorações foram calculadas como o inverso da probabilidade de selecção para uma amostra auto-ponderada dentro de cada distrito, da seguinte maneira:

$$W_{h(p)} \text{ ' } \frac{M_h^{)}}{n_h \times 8} \text{ ' } \frac{M_h^{)}}{m_h} \text{ ,}$$

onde:

$W_{h(p)}$ = factor de expansão para as pequenas explorações seleccionadas no distrito h

M'_h = número total de agregados familiares no marco para o distrito h, estimado das projecções demográficas

m_h = número de agregados familiares com pequenas explorações seleccionadas para o CAP no distrito h

Na realidade, quando se comparou o número de agregados familiares listados (M'_{hi}) com a medida correspondente do recenseamento (M_{hi}) para algumas UPAs, encontrou-se diferenças significativas, provávelmente por causa de inconsistência nas definições da área coberta pela UPA. Dadas as circunstâncias e a informação disponível, é razoável tratar a amostra de pequenas explorações como auto-ponderada dentro de cada distrito para o CAP.

No caso das médias explorações, todas foram incluídas no CAP dentro das UPAs seleccionadas. Então a probabilidade de selecção seria igual à probabilidade da UPA, que é a primeira componente da probabilidade expressa anteriormente, ou seja:

$$p_{hi(m)} \text{ ' } \frac{n_h \times M_{hi}}{M_h} \text{ ,}$$

onde:

$p_{hi(m)}$ = probabilidade de selecção de médias explorações na i-éssima UPA amostral no estrato (distrito) h

O factor de expansão para as médias explorações ($W_{hi(m)}$) seria o inverso desta probabilidade:

$$W_{hi(m)} \text{ ' } \frac{M_h}{n_h \times M_{hi}}$$

O problema para calcular os factores de expansão para as médias explorações é que não temos

informação boa para as medidas de tamanho M_{hi}; quando se comparou o número de agregados do recenseamento para uma UPA com os dados da listagem, muitas vezes eram inconsistentes. Por esta razão os factores de expansão para as médias explorações usadas para as tabelas do CAP foram calculadas baseadas na média das medidas de tamanho para as UPAs dentro do distrito; todas as médias explorações dentro do distrito receberam o mesmo factor de expansão. Dados estas circunstâncias, com a falta de informação sobre o tamanho das UPAs, este procedimento foi razoável. Entretanto, esta metodologia pode introduzir uma possível tendenciosidade na componente das estimativas das médias explorações. Pode-se ver na fórmula para este factor de expansão para as médias explorações que quando a medida de tamanho M_{hi} aumenta, o factor correspondente diminue. Por exemplo, se uma UPA é duas vezes maior que o tamanho médio, seu factor seria a metade do factor medio para o distrito. Ao mesmo tempo, uma UPA grande tem uma chance maior de conter mais médias explorações.

Para a maioria das culturas e das espécies de animais, a contribuição das médias explorações nas estimativas de totais é relativamente baixa; neste caso, a tendensiodidade nos factores de expansão para as médias explorações não seria séria. Entretanto, no caso de gado bovino, as médias explorações correspondem a quase um quarto da estimativa total do número de cabeças. Por isso, seria interessante estudar um pouco mais o efeito desta tendenciosidade.

Em realidade, o número de médias explorações identificadas dentro de uma UPA amostral depende mais do tamanho da lista de agregados familiares para a UPA. Por isso, um factor alternativo que se pode usar para as médias explorações seria o seguinte:

$$W'_{hi(m)} = \frac{M_h}{n_h \times M'_{hi}}$$

onde:

M'_{hi} = número total de agregados familiares identificados na listagem para a i-éssima UPA amostral no distrito h

Talvez seria possível fazer um pequeno estudo para medir a tendenciosidade no caso de algumas UPAs com varias médias explorações que tem gado bovino. Seria necessário buscar a informação da lista de agregados para estas UPAs, calcular este factor de expansão alternativo, e comparar com o factor correspondente que foi usado para o CAP.

4. Tabulação de Erros Amostrais para os Resultados do CAP 1999-2000

Na publicação dos resultados do CAP, é importante incluir uma secção sobre a exatidão dos dados. Esta secção deve incluir estimativas dos erros amostrais para as características mais importantes do CAP e também uma descrição dos erros não amostrais.

O erro amostral é medido pelo erro padrão, que é a raíz quadrada da variância da estimativa. Usou-se o programa CENVAR para calcular os erros padrão para estimativas da área total de culturas, número total de animais e área total de explorações, para Moçambique, províncias e distritos. O CENVAR é uma componente do pacote IMPS (*Integrated Microcomputer Processing System*), usado para as várias estapas de processamento de dados para censos e inquéritos. O INE usou IMPS para processar os dados do RGPH 1997 e vários inquéritos de agregados. O INE e MADER tem cópias deste pacote e os manuais correspondentes; estes também podem ser baixados pelo Internet (ver www.census.gov).

O programa CENVAR calcula os erros padrão usando um estimador chamado "conglomerados últimos." As fórmulas usadas por CENVAR para o cálculo de variancias estão definidas na Secção 11 sobre a Metodologia de CENVAR para o Cálculo de Variâncias. Este estimador de variancias toma em conta o desenho da amostra estratificada de várias etapas. Os quadros produzidos por CENVAR incluem para cada domínio (categoria de classificação) a estimativa ponderada, o erro padrão, o coeficiente de variação (CV, ou erro padrão relativo), o intervalo de confiança de 95 porcento e o efeito do desenho (DEFF). O efeito do desenho para uma estimativa do inquérito é definido como a variancia da estimativa baseado no desenho da amostra complexa, dividida pela variancia correspondente de uma amostra aleatória simples do mesmo tamanho; é uma medida da eficiência relativa do desenho da amostra.

No caso da análise de CENVAR para a pecuária, os dados do CAP incluem as grandes explorações, que tem um factor de expansão de 1, dado a contribuição importante destas explorações para certos animais e domínios. O componente da variância destas grandes explorações é 0, dado que são auto-representadas (incluídas na amostra com uma probabilidade de 1), mas sua inclusão na análise têm o efeito de diminuir os CVs.

Os quadros de CENVAR para o CAP estão apresentados no Anexo I. No caso de culturas, sómente os resultados de CENVAR a nível de distrito para milho são incluídos no anexo. Os resultados distritais para os outros cultivos foram entregues em forma de ficheiro em disquete.

Pode-se ver nos quadros de CENVAR no Anexo I que à nível nacional, os resultados são precisos para a maioria de culturas e animais. No caso do milho, que é a cultura mais predominante, o coeficiente de variação (CV, ou erro padrão relativo) é sómente 1.1 porcento. As outras culturas importantes também tem boa precisão. Entretanto, para algumas culturas mais raras, a precisão sofre por causa das poucas observações que aparecem na amostra do CAP. Por exemplo, o CV para beringela é 61.6 porcento (baseado em sómente 10 observações); para pimenta, 35.3 porcento (81 casos); para gengibre 37.2 porcento (17 casos); para soja, 34.4 porcento (67 casos); e para cenoura, 28.9 porcento (49 casos). No caso dos dados do CAP para a pecuária, os resultados também são precisos para a maioria dos animais. No caso de galinhas, o mais predominante, o CV é de 1.9 porcento. Como se espera, os CVs são mais altos para os

animais mais raros, como os gansos (43.7 porcento, baseado em 52 observações).

A estimativa de área total de explorações é muito preciso, com uma CV de sómente 0.8 porcento ao nível nacional. Mesmo ao nível provincial e distrital as estimativas de área total têm boa precisão.

A nível provincial, os resultados do CAP para as culturas e animais predominantes também são geralmente precisos. Por exemplo, os resultados para milho e galinhas tem boa precisão em todas a províncias. Entretanto, dada as diferênças agro-precuárias por província, algumas culturas e animais tem um CV relativamente alto para algumas províncias. Por isso é importante referir aos resultados de CENVAR no Anexo I para ver a importância relativa das culturas e animais para cada província.

Como se espera, os resultados do CAP a nível distrital são mais sujeitos a erros amostrais. Para os distritos maiores, a precisão é razoável para as culturas e animais mais predominantes. Entretanto, a maioria dos CVs são relativamente altas para os distritos. Por isso os resultados a nível distrital devem ser usadas com cuidado, mas podem servir como indicadores gerais da importância relativa das culturas e animais em cada distrito.

5. Planos Preliminares para a Amostra do TIA 2002

O desenho da amostra preliminar para o TIA 2002 foi elaborado por Chris Hill. A amostra do CAP vai servir de base de amostragem (como uma amostra mãe) para seleccionar uma sub-amostra de UPAs para o TIA. Este desenho inclui três etapas de amostragem. Para melhorar a eficiência da amostra, a estratificação foi baseada em 15 zonas agro-ecológicas definidas para Moçambique.

Cada distrito foi classificado numa destas zonas agro-ecológicas. No caso de um distrito que inclui mais de uma zona, foi assignado à zona predominante em referência à área. Outra alternativa que se poderia considerar para a classificação de distritos mistos seria a predominância de zonas por população. As zonas agro-ecológicas também podem servir como domínios de análise. Entretanto, no caso de distritos que incluem mais de uma zona, a análise pode ser baseado numa post-estratificação dos segmentos (UPAs do CAP) seleccionadas por zona.

Dentro de cada província, um estrato separado foi criado para cada zona agro-ecológica que aparece na província. Assim, foram definidos um total de 26 estratos, apresentados no Quadro 1.

Para reduzir os custos da operação de campo, na primeira etapa foram seleccionados 66 distritos. Dentro de cada estrato, os distritos foram seleccionados com probabilidade proporcional ao tamanho, usando como medida de tamanho o número total de agregados familiares no distrito baseado nos dados do RGPH 1997. O número de distritos seleccionados em cada estrato é apresentado no Quadro 1. No caso de quatro estratos (0205, 0206, 0311, 0614), todos os distritos foram seleccionados para o TIA, porque cada um destes estratos sómente tinha dois

distritos. Neste casos, cada distrito pode ser considerada como um estrato separado.

No segunda etapa de amostragem, 8 UPAs amostrais do CAP são seleccionadas dentro de cada distrito na amostra do TIA. Isto resulta numa amostra aproximadamente auto-ponderada dentro de cada estrato para as pequenas explorações, dado que os distritos foram seleccionados com probabilidade proporcional ao tamanho na primeira etapa. Na realidade, as UPAs do CAP seriam unidades secundárias de amostragem (USAs) para o TIA. Para evitar confusão, o relatório vai referir as estas unidades como "segmentos."

Na terceira etapa de amostragem, uma nova lista de agregados familiares dentro da UPA vai ser usada para seleccionar 8 pequenas explorações. Como no CAP, as médias explorações vão ser incluídas com certeza na terceira etapa da amostra.

Quadro 1. Atribuição da Amostra Proposta para o TIA por Estrato: Distritos, Segmentos e Pequenas Explorações

Província	Estrato	No. Distritos CAP	No. Distritos TIA	No. Segmentos TIA	No. Peq. Explor. TIA
Niassa	0101	5	2	16	128
	0106	11	2	16	128
Cabo Delgado	0205	2	2	16	128
	0206	2	2	16	128
	0210	13	4	32	256
Nampula	0306	3	2	16	128
	0310	15	6	48	384
	0311	2	2	16	128
Zambézia	0404	3	2	16	128
	0407	5	4	32	256
	0411	9	4	32	256
Tete	0502	3	2	16	128
	0508	3	2	16	128
	0512	7	2	16	128
Manica	0603	5	2	16	128
	0609	3	2	16	128
	0614	2	2	16	128
Sofala	0709	4	2	16	128
	0711	5	2	16	128
	0712	3	2	16	128
Inhambane	0813	11	4	32	256
	0814	3	2	16	128
Gaza	0913	3	2	16	128
	0914	5	2	16	128
	0915	4	2	16	128
Maputo-Provincia	1014	7	4	32	256
Moçambique		138	66	528	4,224

6. Análisis dos Erros Padrão Simulados para o TIA 2002

Dado a importância do TIA para proporcionar estatísticas agro-pecuários anuais e o grande investimento correspondente do Governo de Mozambique e doadores, é importante determinar o nível aproximado de precisão que se pode esperar para os resultados mais importantes do inquérito, como também medir a eficiência da amostragem. Para este propósito foi elaborada uma simulação dos erros padrão, usando os dados do CAP para os segmentos seleccionados para o TIA. O pacote CENVAR voi usado para fazer esta análise dos erros padrão e efeitos do desenho, tomando em conta o desenho proposto para o TIA 2002.

Para refletir bem o desenho da amostra do TIA na análise de CENVAR, foi necessário gerar códigos apropriados para identificar os estratos e UPAs na amostra. Na maioria das zonas agro-ecológicas, o estrato é a zona, e a UPA é a provincia. Mas no caso das quatro zonas onde todos os distritos foram seleccionados, o estrato é o distrito, e a UPA é o segmento. Dada a alta taxa de amostragem para os distritos na primeira etapa, foi necessário de incluir na análise de CENVAR um ajuste por população finita, baseada na probabilidade média da primeira etapa dentro de cada estrato. As fórmulas usadas por CENVAR estão apresentadas na Secção 11.

Os resultados de CENVAR para este estudo de erros padrão simulados para o TIA 2002 estão apresentados no Anexo 2. Para esta análise, dados do CAP não estavam disponíveis para 13 dos segmentos seleccionados para o TIA, que representa um pouco mais de 2 porcento da amostra de 528 segmentos; 9 destes segmentos amostrais que faltam dados são do estrato 0411 em Zambézia, de um distrito não coberto pelo CAP por causa das inundações. Entretanto, em geral os resultados deste estudo de simulação devem medir bem os erros padrão e efeitos do desenho aproximados baseado no desenho do TIA.

Os resultados de CENVAR para esta análise de simulação estão apresentados no Anexo II. Pode-se ver nestes quadros que a nível nacional só nove culturas tem um CV menor que 10 porcento (que se pode considerar boa precisão): milho, mandioca, pepino, mapira, batata doce, abóbora, feijão nhemba, feijão jugo e amendoím. Outros três cultivos tem um CV entre 10 e 15 porcento (que se pode considerar precisão razoável): arroz, mexoeira e feijão boer.

A média dos efeitos do desenho para estas doze culturas que têm melhor precisão. As culturas com efeitos do desenho mais altos são feijão manteiga (16.92, baseado em 641 observações) e melancia (13.11, com 465 observações). No caso de culturas com poucas observações, alguns efeitos do desenho são menores de 1, mas estes resultados não são confiáveis.

No caso dos dados de pecuaria para a amostra do TIA, os seguintes animais tem uma boa precisão a nível nacional: galinhas, patos, caprinos e suínos. A precisão também é razoável para bovinos (com CV de 11.3 porcento). Os efeitos do desenho para pecuária são artificialmente baixos, provávelmente por causa da inclusão de grandes explorações no ficheiro de dados.

Para o nível provincial, a única cultura com dados de boa precisão para todas a províncias é o milho. A precisão das outras culturas varia por província. No caso de pecuária, sómente os

resultados para galinhas são precisos para todas as províncias. Algumas províncias também tem uma precisão razoável para certos outros animais, mas em geral os CVs são altos.

7. Resultados Comparativos de Um Desenho Alternativo para TIA 2002

Dado os efeitos de desenho relativamente altos que resultam da selecção de províncias na primeira etapa de amostragem para o TIA 2002, seria interessante de comparar estes resultados com aquêles que resultariam de uma amostra bi-etápica. Para fazer este novo estudo de simulação, a base de amostragem do CAP foi usado para seleccionar uma amostra de segmentos (neste caso, UPAs do CAP) dentro de cada estrato (província por zona agro-ecológica). Para manter uma comparabilidade directa com os resultados do desenho preliminar do TIA, usou-se a mesma estratificação e atribuição de segmentos e explorações para a amostra alternativa.

Para uma amostra de duas etapas, também seria ideal seleccionar uma amostra de pequenas explorações aproximadamente auto-ponderada dentro de cada estrato. Neste caso, não se pode seleccionar as UPAs do CAP com probabilidades iguais dentro de cada estrato, porque a taxa de amostragem no CAP varia por distrito (isto é, o estrato para o CAP). Examinando os factores de expansão para o CAP, se pode ver que variam de um distrito para outro baseado nas diferênças das razões M_h/n_h (quer dizer, o número total de agregados familiares no distrito dividido pelo número de UPAs seleccionadas no distrito). No caso de uma atribuição proporcional da amostra, esta razão seria constante. Para controlar esta variabilidade nas taxas de amostragem por distrito no CAP, as UPAs para a amostra alternativa de TIA foram seleccionadas proporcionalmente ao factor M_h/n_h para o distrito.

Um ficheiro Excel foi usado para fazer esta selecção de UPAs sistemáticamente com probabilidade proporcional a M_h/n_h dentro de cada estrato (província por zona agro-ecológica). Este ficheiro contem informação para cada UPA na amostra do CAP. Uma amostra de 528 UPAs foi seleccionada usando o desenho alternativo para o TIA; a atribuição das UPAs amostrais aos estratos foi o mesmo daquela apresentada no Quadro 1.

Os dados do CAP para estas 528 UPAs foram usados para esta análise de erros padrão simulados baseado no desenho alternativo de TIA. Por falta de tempo, esta análise foi limitada aos dados para culturas. Neste análise de CENVAR não usou-se um factor de ajuste por população finita, como no caso da aplicação para o desenho preliminar do TIA; isto pode resultar em resultados para os erros padrão um pouquinho elevados, mas assumimos que as taxas da primeira etapa de amostragem é geralmente menos de 5 porcento. Os resultados deste análise simulado de CENVAR estão apresentados no Anexo III. Pode-se ver que em geral os CVs e efeitos do desenho baixaram considerávelmente comparados com os resultados correspondentes da amostra de três etapas. Por exemplo, no caso das doze culturas com melhor precisão, o efeito de desenho médio baixou de 4.1 a 2.9. A razão entre estes dois efeitos (2.9/4.1), 0.71, quer dizer que usando uma amostra bi-etápica se pode baixar o tamanho da amostra por 29 porcento para obter o mesmo nível de precisão que vai-se obter do desenho do TIA baseado na selecção de 66 distritos na primeira etapa. Também se pode notar no Anexo III que o número de culturas com CV menor a 15 porcento a nível nacional subiu de 12 a 17. As seguintes cinco culturas também

teriam uma precisão boa ou razoável: feijão manteiga, quiabo, melancia, algodão e cana de açúcar. A nível de cultura por província, encontramos 16 casos em que a CV baixa a menos de 15 porcento quando se vai de três a duas etapas. Para culturas com CVs mais altas, encontramos também alguns casos em que o CV é mais baixo usando três etapas, dado que usamos amostras diferentes; em geral estes casos podem ser considerados menos estáveis por causa do pequeno número de observações.

A amostra alternativa de duas etapas de amostragem para o TIA inclui UPAs seleccionadas em 132 distritos, que são quase todos, e isto implicaria um maior custo para o trabalho de campo. A razão de seleccionar distritos na primeira etapa de amostragem era de reduzir os custos de transporte durante o inquérito, dado que vai concentrar a amostra na metade dos distritos. Entretanto, se pode ver desta análise simulada da eficiência de amostragem que este desenho também implica um custo mais alto em termos de tamanho de amostra requerida.

Dado a grande diferênça na eficiencia do desenho entre uma amostra de duas e três etapas, o consultor recomenda examinar bem as implicações de custos de cada alternativa. O INE tem muita experiência com inquéritos nacionais de agregados familiares usando amostra bi-etápica, com as UPAs dispersas entre os distritos dentro de cada província, incluíndo o Inquérito de Agregados Familiares (IAF) e o Inquérito de Demografia e Saúde. Pode-se aproveitar desta experiência para determinar as implicações de custo e logística operativa de usar uma amostra dispersa entre os distritos. A amostra alternativa de 528 UPAs está identificada no ficheiro Excel chamado tia2samp.xls. Seria interessante examinar a dispersão destas UPAs num mapa. Por exemplo, em vez de usar o distrito como base de operações do campo, se pode dividir as UPAs amostrais em grupos operativos baseado nos mapas, vias de transporte, e considerações operativas.

Dada a precisão limitada dos resultados baseado no desenho preliminar do TIA apresentados no Anexo II, outra maneira de interpretar esta análise é que é possível melhorar a precisão dos resultados do TIA usando uma amostra de duas etapas do mesmo tamanho planejado. Assim se pode obter estimativas com uma precisão similar aos resultados apresentados no Anexo III. Isto depende dos recursos disponíveis para o inquérito.

Outra vantagem de uma amostra bi-etápica é que seria possível de estratificar as UPAs individuais do CAP em zonas agro-ecológicas. Isto evitaria o problema de distritos que comtem mais de uma zona, e melhoraria a eficiência da amostra e da análise.

8. Considerações para a Estratificação de Explorações por Tamanho

Na estimação de totais de área e produção de culturas, e número total de animais, a contribuição relativa das grandes explorações é importante. Por isso o consultor recomenda manter uma base separada de grandes explorações que podem ser incluídas na amostra de TIA com certeza.

Pela mesmo razão, as médias explorações devem ser seleccionadas com uma taxa de amostragem maior que as pequenas, similar à metodologia do CAP. Uma sugestão discutida com Chris Hill é

a possibilidade de ajustar o tamanho mínimo a 5 hectares na definição de médias explorações. Outros critérios podem ser estabelecidos para as explorações pecuárias, baseada no número total de animais por tipo. Se todas estas explorações encontradas dentro dos segmentos seleccionados são incluídas na amostra do TIA, isso vai melhorar a precisão das estimativas de totais. Ao mesmo tempo, esta alternativa implica um pequeno aumento no tamanho de amostra e custo do inquérito. Primeiro seria necessário estimar o número de médias explorações que podem aparecer nos segmentos seleccionados, baseado nos dados do CAP. De acordo com a distribuição ponderada do CAP, as explorações com 5 a 9.9 hectares representam 1.61 porcento do total, e as explorações com 10 ou mais hectares representam sómente 0.14 porcento; neste caso, o total com 5 ou mais hectares seria 1.75 porcento. Isto que dizer que dentro de um segmento amostral de 120 agregados familiares uma média de duas explorações vão ter 5 ou mais hectares. Neste caso o número total de explorações na amostra dentro da UPA seria 10, incluíndo as 8 explorações pequenas.

Se os recursos para o TIA não são suficientes para incluir todas as explorações com 5 ou mais hectares no inquérito, outra alternativa que se pode considerar seria de identificar todas as explorações dentro do segmento com 5.0 a 9.9 hectares, e aquelas com 10 ou mais hectares. Se o número de explorações com 5 a 9.9 hectares é menos de 4, deve-se incluir todas as explorações com 5 ou mais hectares no TIA. Se o número de explorações com 5 a 9.9 hectares é de 4 ou mais, se pode seleccionar 50 porcento; as explorações com 10 ou mais hectares continuariam a ser incluídas com certeza dentro do segmento amostral.

9. Considerações Amostrais para a Comparibilidade dos Resultados Anuais de TIA

Dado que o TIA vai ser um inquérito anual, também é importante considerar como a amostra vai ser seleccionada nos próximos anos. Um objectivo chave do TIA é de medir as diferênças anuais na produção agro-pecuária e nos indicadores sócio-económicos. Por esta razão seria vantajoso reter (sobrepôr) uma parte da amostra (por exemplo, 75 porcento) de um ano ao outro. Pode ser difícil manter as mesmas explorações na amostra cada ano, mas pelo menos as mesmas UPAs amostrais podem ser usadas para maximizar a correlação entre as amostras anuais. Se um desenho de três etapas de amostragem é usado para o TIA, não se deve seleccionar distritos diferentes cada ano, porque isto vai minimizar a correlação entre as amostras necessária para melhorar a fiabilidade das estimativas de mudança anual.

10. Considerações para a Operação de Listagem e a Mantenção da Base de Amostragem

No caso do CAP, a selecção de agregados familiares dentro das UPAs amostrais foi baseada numa listagem que em geral foi feita pelos inquiridores de casa em casa. Entretanto, em alguns casos que escaparam ao controle dos gestores provinciais do CAP, se usou uma lista de famílias fornecida pelos chefes das UPAs. Em geral, um problema com este tipo de lista é que é difícil controlar a qualidade da cobertura. Por exemplo, é possível que alguns agregados mais marginais ou isolados não sejam incluídos na lista. Também é difícil determinar a medida de tamanho real de um segmento para calcular os factores de expansão, como no caso das médias explorações no CAP.

Por esta razão o consultor recomenda que dentro de cada segmento seleccionado para o TIA, se faça uma operação de listagem bem controlada, em que o inquiridor vai de casa em casa para obter a lista completa de agregados familiares. Para este propósito podem aproveitar da experiência do INE com operações de listagem que fizeram para vários inquéritos de agregados familiares. Uma referência boa seria o manual de listagem que usaram para o último Inquérito de Demografia e Saúde. Também é importante de manter informação sobre o número total de agregados familiares listados em cada UPA num ficheiro que pode ser usado para o cálculo de factores de expansão e para controlar a amostra de UPAs através do tempo.

11. Metodologia de CENVAR para o Cálculo de Variâncias

O estimador da variancia deve tomar em conta os diferentes aspectos do desenho de amostragem, como a estratificação e a conglomeração. O programa CENVAR poder ser usado para calcular as variancias para estimativas de um inquérito baseado numa amostra estratificada de várias etapas, como o CAP e o TIA 2002. O programa CENVAR é amigável e baseado em menus. Usa um dicionário definido baseado no programa DATADICT de IMPS. O CENVAR pode ser usado para calcular os erros padrão para totais, médias, proporções e outros tipos de razões. O CENVAR pode gerar estimativas de sub-população para cada categoria de uma variável de classificação, e estas variáveis podem ser cruzadas. Para cada categoria, CENVAR calcula a estimativa, erro padrão, CV, intervalo de confiança de 95 porcento e o efeito do desenho (DEFF). Exemplos do tipo de quadro produzido por CENVAR estão apresentados nos Anexos I, II e III, com os resultados do CAP e os resultados simulados para o TIA 2002. Estes são exemplos de totais para área de explorações, área em culturas e número de animais.

Para calcular estimativas de erros padrão usando CENVAR, geralmente é necessário gerar um novo ficheiro de dados em formato de texto dos dados originais do inquérito. Dado que o programa CENVAR só pode aceitar um tipo de registro, é necessário gerar um ficheiro de dados para CENVAR com um registro para cada unidade de análise. Por exemplo, no caso de estimativas por exploração, como total de área, foi necessário gerar um registro para cada agregado na amostra. Para estimativas por cultura, como a área semeada por cultura, o ficheiro de dados para CENVAR deve ter um registro para cada cultura dentro de cada parcela. Cada registro deve ter campos para o código de estrato, conglomerado (ou UPA) e factor de ponderação, como também as variáveis de classificação e análise. Quando a taxa de selecção das UPAs é relativamente alta (em geral, mais de 5 porcento), como no caso do desenho do TIA, também seria necessário incluir a probabilidade da primeira etapa de amostragem no ficheiro de dados para CENVAR.

As variáveis de classificação são usados para produzir estimativas para todas as categorias respectivas. As variáveis de análise geralmente são variáveis contínuas, como área e número de animais, ou variáveis binárias, que são igual a 1 se a unidade tem a característica de interesse, e 0 se não tem. O programa CENVAR automáticamente gera uma variável chamada INTERCEPT, que é igual a 1 para cada registro. A variável INTERCEPT pode ser usado para obter a estimativa do número total ponderado de unidades (por exemplo, o número total de explorações), ou pode ser usado no denominador de uma razão para obter uma média ou proporção. O ficheiro

de dados do inquérito gerado para CENVAR deve ser ordenado por estrato e conglomerado (UPA).

O programa CENVAR não aceita campos vazios no ficheiro. As variáveis de classificação com falta de dados devem ser substituídas por um código para identificar "não resposta" ou "não aplicável." Os resultados de CENVAR vão incluir estimativas para estas categorias, que podem ser eliminadas dos quadros que vão ser publicadas. Quando os espaços vazios no ficheiro são substituidos por 0 (zero), qualquer variável de análise que está faltando seria tratado como zero, que pode tendenciar para baixo as estimativas de médias. Por exemplo, se um questionário está faltando informação para área, o CENVAR vai considerar que o valor da área para esta exploração é zero. Duas alternativas podem ser consideradas para evitar este tipo de tendência. Uma alternativa seria de eliminar os registros que estão faltando as variáveis de análise. Outra opção seria de introduzir uma variável de classificação adicional, que teria um código de 1 se os dados para a variável de análise (como, por exemplo, área ou número de animais) existem para este registro, e 2 se não existem. Depois as estimativas para a categoria de "dados faltando" (código 2) podem ser eliminados dos quadros com os resultados de CENVAR.

O programa CENVAR usa um estimador de variancia chamado "conglomerados últimos." Para a estimativa de um total, a variancia é calculada por CENVAR usando a seguinte fórmula:

Estimador de Variancia para um Total

$$V(\hat{Y}) = \sum_{h=1}^{L} \left[\frac{(N_h - n_h)}{N_h} \times \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \right],$$

onde:

$\hat{Y}_{hi} = \sum_{j=1}^{m_{hi}} W_{hi}\, y_{hij}$ = total ponderado para a i-éssima UPA amostral dentro do estrato h

m_{hi} = número de unidades de análise (por exemplo, explorações) na amostra para a i-éssima UPA amostral dentro do estrato h

$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$ = total ponderado para o estrato

O factor $\frac{(N_h - n_h)}{N}$ representa o ajuste por população finita que sómente seria incluído quando está especificado no análisis do CENVAR. Seria usado quando a taxa de selecção das UPAs é

relativamente alta, como no caso do TIA onde quase a metade dos distritos foram incluídos na amostra na primeira etapa.

Para a estimativa de uma razão, a variância é calculada por CENVAR usando a seguinte fórmula:

Estimador de Variância para uma Razão

$$V(\hat{R}) \approx \frac{1}{\hat{X}^2}\left[V(\hat{Y}) + \hat{R}^2\, V(\hat{X}) - 2\,\hat{R}\, COV(\hat{X},\hat{Y})\right],$$

onde:

$$COV(\hat{X},\hat{Y}) \approx \sum_{h=1}^{L}\left[\frac{(N_h - n_h)}{N_h}\times\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}\left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h}\right)\left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h}\right)\right]$$

$V(\hat{Y})$ e $V(\hat{X})$ são calculados usando a fórmula para a variância da estimativa de um total, especificada anteriormente.

O consultor apresentou um cursinho práctico de CENVAR para um grupo de 18 estatísticos e economistas do INE e MADER. Esta formação de um dia começou com uma exposição de conceitos básicos de amostragem e estimação de variância. Durante o curso, a maioria dos participantes usaram computadores tipo "lap-top" para seguir vários exemplos usando CENVAR e outros componentes de IMPS: DATADICT para definir o diccionário e QUICKTAB para produzir quadros de frequências e tabelas cruzadas. No final do curso foram apresentados a aplicação e resultados de CENVAR para o CAP. Os participantes foram muito atentos, e aproveitaram bem dos exercícios de CENVAR. Ganharam experiência seguindo todos os passos necessários para produzir uma aplicação de CENVAR.

12. Alternativa de Seleccionar Sub-Amostra de TIA para Indicadores de Rendimento Familiar

Um objectivo do TIA vai ser de medir indicadores relacionados a receitas ou despesas dos agregados familiares. Uma opção que estão considerando é de usar o questionário correspondente numa sub-amostra de TIA. Neste caso devemos calcular o tamanho mínimo da amostra necessária para cada domínio de análise, e determinar o número de domínios. Para determinar o tamanho aproximado da amostra, podemos usar os dados de despesas por agregado familiar do IAF 1996/97. O Quadro 2 apresenta os resultados de CENVAR para estimativas de médias de despesas por agregado para a parte rural de cada província.

Quadro 2. Erros Padrão Calculados por CENVAR para Estimativas de Médias de Despesas por Agregado para Residência Rural de Cada Província, IAF 1996/97

Categoria	Estimativa	Erro Padrão	C.V. (%)	Intervalo de Confiança, 95% Inferior	Intervalo de Confiança, 95% Superior	DEFF	Número de Observações
PROVÍNCIA POR ÁREA							
Niassa	505,734	44,582	8.82	418,353	593,115	3.94	504
Cabo Delgado	552,929	33,111	5.99	488,031	617,827	4.11	646
Nampula	445,739	31,994	7.18	383,031	508,448	7.67	719
Zambêzia	518,457	19,445	3.75	480,346	556,569	6.60	790
Tete	485,981	33,068	6.80	421,168	550,795	4.61	504
Manica	786,708	44,788	5.69	698,923	874,493	1.88	485
Sofala	348,508	26,713	7.66	296,151	400,866	3.59	504
Inhambane	610,953	36,529	5.98	539,356	682,550	2.26	601
Gaza	1,001,189	73,177	7.31	857,761	1,144,616	3.01	563
Maputo-Província	1,869,109	132,180	7.07	1,610,036	2,128,183	1.34	282
Moçambique Rural	558,666	12,376	2.22	534,409	582,922	3.98	5,745

Neste caso usamos a parte rural de cada província como exemplo de domínios de análise. Supondo que aceitaríamos um erro relativo de 10 porcento (isto é, o intervalo de confiança de 95 porcento seria igual à estimativa ± 10%), seria necessário obter uma CV de 5 porcento. Neste caso podemos usar o erro padrão e o número de agregados familiares na amostra para o IAF 1996/97 para estimar o tamanho da amostra requerida para obter uma CV de 5 porcento em cada domínio. A razão entre a variância medida por CENVAR com os dados do IAF 1997 para a média de despesas por agregado para o domínio e a nova variância correspondente requerida para o TIA neste domínio pode ser expressada da seguinte maneira:

$$\frac{se^2{}_1(\bar{x}_d)}{se^2{}_2(\bar{x}_d)} \quad , \quad \frac{\frac{\sigma_{xd}{}^2}{m_1}\times DEFF_1}{\frac{\sigma_{xd}{}^2}{m_2}\times DEFF_2} \quad ,$$

onde:

$se_1(\bar{x}_d)$= erro padrão para a estimativa de média de despesas por agregado para o domínio d baseado no desenho do IAF 1996/97, calculado por CENVAR

$se_2(\bar{x}_d)$ = erro padrão desejada para a estimativa de média de despesas por agregado para o domínio d baseado no desenho do TIA

$\sigma_{xd}{}^2$= variância da população (quadrado do desvio padrão) para a variável despesas total por agregado, para o estrato h

$DEFF_1$ = efeito do desenho para a estimativa de média de despesas por agregado para o

domínio, do IAF 1996/97

m_1 = número de agregados na amostra do IAF 1996/97 no domínio d

m_2 = número de agregados requiridos na amostra do TIA 2002 no domínio d, para obter o nível de precisão desejado

$DEFF_2$ = efeito do desenho para a estimativa de média de despesas por agregado para o TIA 2002

Supondo que os efeitos do desenho para os dois inquéritos são similares, se pode simplicar esta expressão para calcular o tamanho de amostra requerida para o domínio no TIA assim:

$$m_2 \text{ ' } \frac{se^2{}_1(\bar{x}_d)}{se^2{}_2(\bar{x}_d)} \times m_1$$

Esta fórmula foi usada para calcular o tamanho de amostra mínimo requerido para o domínio rural de cada província para obter uma CV de 5 porcento; estes resultados estão apresentados no Quadro 3.

Quadro 3. Tamanho de Amostra Calculada para Obter uma CV de 5 Porcento Para Estimativas de Médias de Despesas por Agregado Familiar Dentro de Cada Domínio Rural por Província

Província	Estimativa	se_2	CV	m_2
Niassa	505,734	25,287	5.0	1,567
Cabo Delgado	552,929	27,646	5.0	927
Nampula	445,739	22,287	5.0	1,482
Zambezia	518,457	25,923	5.0	445
Tete	485,981	24,299	5.0	933
Manica	786,708	39,335	5.0	629
Sofala	348,508	17,425	5.0	1,184
Inhambane	610,953	30,548	5.0	859
Gaza	1,001,189	50,059	5.0	1,203
Maputo-Provincia	839,263	41,963	5.0	523
Moçambique Rural	558,666	27,933	5.0	1,128

Pode-se ver no Quadro 3 que o tamanho de amostra requerido varia por província, dado as diferênças na variabilidade da característica despesas por província. A média dos tamanhos de

amostra m_2 por domínio (província) rural é 975. Por isso, se pode considerar que em geral seria necessário seleccionar uma média de 1,000 agregados familiares por domínio para obter este nível de precisão. No caso do domínio rural que inclui todas as províncias de Moçambique, o tamanho mínimo da amostra requerida para obter uma CV de 5.0 porcento seria 1,128 agregados familiares. Para domínios rurais que são formados por um grupo de províncias, se pode calcular a média dos valores m_2 para as províncias correspondentes. Uma sub-amostra de 50 porcento para o TIA corresponde a uma amostra de aproximadamente 2,100 agregados familiares. Neste caso, estaríamos limitados a dois domínios de análise que poderíamos comparar com confiança.

Caso seja possível ajustar o nível de erro relativo aceitável a 12 porcento, se pode baixar a média do tamanho de amostra necessário para cada domínio a aproximadamente 677 agregados familiares. Neste caso, uma sub-amostra de 50 porcento permitiria até três domínios de análise. Para o domínio rural que abrange todas as províncias de Moçambique, o tamanho de amostra requerido para uma CV de 6 porcento baixaria a 783 agregados familiares.