ELABORAÇÃO E TESTE DE PROGRAMAS COMPUTACIONAIS PARA CORREÇÃO AUTOMÁTICA DE DADOS DE LEVANTAMENTOS AGRÍCOLAS[1]

José Roberto Vicente[2]
Lilian Cristina Anefalos[3]

RESUMO

O objetivo deste trabalho foi elaborar e testar programas computacionais para correção automática de dados do levantamento objetivo das safras agrícolas do Estado de São Paulo, Brasil. Foram desenvolvidos oito módulos empregando rotinas e procedimentos do SAS (*Statistical Analysis Software*), procurando corrigir os erros de preenchimento mais comuns: respostas incompletas, perguntas não compreendidas, erros de cálculos, erros de linhas e respostas fora dos limites. Foram discutidas com detalhes as rotinas de correção de amendoim das águas, feijão das águas e milho (módulo 1) e as de correção de produção de leite (módulo 8). Os resultados destes programas foram comparados com os provenientes de correção manual e, na grande maioria dos casos, não diferiram estatisticamente destes últimos.

Palavras-chave: depuração de dados, correção automática, levantamentos agrícolas.

DEVELOPMENT AND SIMULATION OF SOFTWARE ROUTINES FOR AUTOMATIC DATA CORRECTION OF AGRICULTURAL SURVEYS

SUMMARY

The objective of this paper was to develop and simulate software routines for the automatic correction of the data of the objective survey of agricultural crops of São Paulo State, Brazil. Eight modules have been elaborated by using SAS (Statistical Analysis Software) routines and procedures to correct common errors of fulfilling: incomplete answers, non understanding of questions, calculation errors, line errors and answers out of limits. The routines of rain peanuts, rain beans and corn (module 1) as well as milk production (module 8) were widely discussed. The results were contrasted with manual correction and their great majority was considered statistically similar to it.

Key-words: data depuration, automatic correction, agricultural surveys.

1 - INTRODUÇÃO

Conhecer a magnitude das safras agrícolas interessa tanto ao poder público como a vários outros setores da economia, devido às influências sobre o abastecimento interno, as exportações e a estabilidade dos preços. Por esse motivo, grandes esforços são efetuados, principalmente pelo poder público, para elaborar e divulgar previsões e estimativas de safras. Em São Paulo, o Instituto de Economia Agrícola (IEA) e a Coordenadoria de Assistência Técnica Integral (CATI) efetuam levantamentos cinco vezes por ano, com questionários em nível de município (levantamento subjetivo, não probabilístico) e em nível de imóvel rural, com amostra probabilística duplamente estratificada (por região e por tamanho de imóvel), atualmente constituída de 3.622 elementos (levantamento objetivo)[4]. O levantamento objetivo, apesar dos custos mais elevados, apresenta diversas vantagens sobre o subjetivo[5], especialmente com respeito à qualidade e profundidade das informações obtidas, e os retornos dele provenientes aparentam ser compensadores[6].

Os questionários dos dois levantamentos são elaborados no IEA e enviados à CATI, que é encarregada do preenchimento dos mesmos. No caso do subjetivo, os responsáveis pelas Casas de Agricultura

fornecem os dados dos municípios, enquanto os do objetivo são coletados junto aos produtores. Após preenchidos, eles são remetidos de volta ao IEA, onde são digitados, conferidos e depurados, visando, principalmente, publicar previsões e estimativas de safras de uma grande variedade de produtos da agricultura paulista.

Para a maioria das culturas, a fonte de dados é o levantamento subjetivo; o objetivo fornece informações somente sobre as principais lavouras do Estado[7]; em outras partes dos questionários há questões sobre pecuária (corte e leite), suinocultura, utilização da terra, uso de insumos e técnicas agrícolas, demografia e mão-de-obra, etc., distribuídas pelos diversos levantamentos efetuados durante o ano.

Uma das partes mais importantes do processo de elaboração de previsões de safras é a fase de depuração de dados. É também monótona, demorada e repetitiva, pouco atraente para pesquisadores, cuja função é de natureza mais exploratória. A elevada rotatividade dos auxiliares de pesquisa vem impedindo que se consiga um quadro suficientemente experiente para responder por essas atividades, que continuam sendo executadas pelos profissionais mais treinados das seções responsáveis pelos levantamentos. A própria essência rotineira dessa fase, em que milhares de operações matemáticas simples são efetuadas quase manualmente (com auxílio somente de calculadoras), faz com que exista um número inestimável de erros em potencial, cometidos na etapa cuja finalidade precípua é a de eliminá-los.

Com a ampliação dos recursos de informática do IEA, diversas fases do processo, que eram executadas em outras Instituições, puderam ser internalizadas e, juntamente com as que já eram efetuadas no Instituto, aperfeiçoadas. Para o levantamento subjetivo, por exemplo, foi elaborado um sistema de processamento que facilitou enormemente a fase de depuração de dados, indicando erros e efetuando automaticamente inclusões de municípios não respondentes, com base nos questionários anteriores, outrora executados manualmente.

Já no levantamento objetivo, maior e computacionalmente mais complexo, apesar de diversas reformulações e aperfeiçoamentos, esse processo pouco evoluiu, permanecendo basicamente da mesma maneira há vinte anos. Os únicos subsídios praticamente incorporados, desde que o método proposto por PINO & OSSIO (1975) passou a ser empregado regularmente, foram planilhas destinadas a diminuir erros de transcrição de correções já efetuadas e o cálculo de relações médias em nível de Estado. Ainda hoje, o processo consiste no estabelecimento de relações que devem ser obedecidas, emissão de listagens de questionários que não atendem às condições, correção manual dos dados e posterior envio para digitação.

Em princípio, a maioria das atividades de correção dos questionários é formada por uma seqüência de operações lógicas pré-determinadas. É factível, portanto, o desenvolvimento de programas computacionais específicos para efetuá-las.

O objetivo deste estudo foi o de elaborar e testar programas capazes de corrigir dados do levantamento objetivo, com a finalidade de agilizar e aperfeiçoar o procedimento de depuração, e minimizar os erros e falhas inerentes ao sistema até então adotado.

2 - METODOLOGIA

A maneira provavelmente mais adequada de eliminar erros de questionários seria o retorno destes aos imóveis rurais informantes, tão logo algum problema fosse detectado para que as informações pudessem ser rigorosamente acertadas[8]. Entretanto, devido ao grande número de imóveis levantados (atualmente 3.622) e à necessidade premente de divulgação imediata das previsões de safras, tal procedimento não é adotado. Inicialmente, a etapa de depuração de dados do levantamento objetivo era efetuada por inspeção visual dos questionários, procurando-se corrigir erros grosseiros de preenchimento. Posteriormente, PINO & OSSIO (1975) desenvolveram um método baseado no cálculo de relações simples que deveriam ser respeitadas, passando-se a trabalhar com listagens emitidas por computador, solucionando apenas casos de valores fora de limites previamente determinados; esse é, em essência, o sistema vigente até hoje. As correções efetuadas baseiam-se na hipótese de que os erros ocorrem de forma aleatória[9]. GRILICHES (1986) analisa as

implicações, para diversos tipos de estimadores, de casos em que essa pressuposição não pode ser aplicada a dados perdidos (*missing values*).

PINO (1986) descreveu a teoria subjacente a diversos procedimentos de detecção e correção de dados. Esse autor descreve a técnica de consistência interna a partir de um resultado simples de probabilidade: sendo X uma variável aleatória, tal que

$$E|X| < \infty \qquad (1)$$

então, dado $0 <$ 1, existem números reais a e b, tais que

$$P(X \notin [a,b]) < \varepsilon \qquad (2)$$

O que significa a possibilidade da construção de um intervalo finito, tal que seja ínfima a probabilidade da variável ter valores fora do mesmo, os quais passam a ser considerados errados ou *outliers*.

PINO (1986) definiu também o valor de um teste por

$$T(X_e) = \frac{f(X_e)}{g(X_e)} \qquad (3)$$

com

$$T(X_e) \in [L_i, L_s], \qquad (4)$$

onde L_i e L_s são números reais ($L_i \leq L_s$), X_e é um vetor aleatório mx1, constituído de m variáveis do levantamento, f e g são funções reais de X_e e T é um teste de consistência interna se f e g tiverem significado prático. Esse autor apresenta quatro possibilidades de erros de consistência interna:

a) $T(X_e) < L_i$;

b) $T(X_e) > L_s$;

c) $L_i = L_s = L$ e $T \neq L$; e

d) $g(X_e) = 0$ e $f(X_e) \neq 0$ ou vice-versa.

A maioria das correções efetuadas nos dados do questionário objetivo enquadra-se no primeiro exemplo, descrito por aquele autor, para aplicações em levantamentos agrícolas: se a produtividade de um imóvel estiver fora de limites pré-definidos, a produção corrigida pelo modelo será dada pela área multiplicada pela produtividade média[10].

Os tipos de erros não amostrais que ocorrem no levantamento objetivo foram classificados por PINO & CASER (1984b) em:

a) resposta incompleta;

b) falta de soma;

c) pergunta não compreendida;

d) erro de cálculo;

e) erro de linha; e

f) resposta fora dos limites.

Com as alterações inseridas nos questionários desde então, o erro tipo B não mais ocorre nos itens referentes à previsão de safras e produção de leite, que foram tratados neste trabalho. Da mesma forma, o erro tipo D pode, para esses mesmos itens, ser enquadrado como tipo F.

As atividades agrícolas existentes no objetivo foram, para efeito da elaboração dos programas, divididas em oito módulos:

1) amendoim das águas, feijão das águas e milho;

2) algodão, milho safrinha e soja safrinha;

3) arroz;

4) feijão de inverno;

5) soja;

6) laranja;

7) café; e

8) produção de leite.

Dentro de cada módulo, as diferenças de procedimentos utilizados foram consideradas menores do que entre cada um deles. Devido à grande massa de dados e à limitação de *hardware*, bem como pela facilidade operacional, optou-se por processar cada produto separadamente dentro do módulo correspondente.

Foi realizada uma simulação inicial, a partir de arquivos contendo as diversas possibilidades de erros, servindo como teste inicial para os programas desenvolvidos no SAS (*Statistical Analysis Software*). Em seguida, utilizou-se o arquivo de dados do levantamento objetivo de setembro de 1994, com os resultados sendo processados e comparados com os provenientes do método de correção manual, através de testes *t*

(de Student):

$$t = \frac{Ca - Cm}{s(Ca)} \qquad (5)$$

onde Ca representa estimativa obtida a partir da correção automática, Cm estimativa proveniente da correção manual e s(Ca) o desvio-padrão de Ca (GOMES, 1970). Esses testes, que para serem estritamente válidos pressupõem normalidade dos dados, foram considerados como significativos nos casos em que os valores de *t* obtidos foram superiores a um[11].

3 - RESULTADOS E DISCUSSÃO

Neste item são apresentados e discutidos os conceitos utilizados em cada módulo de correção. Em seguida, são comparados os resultados dos dois processamentos de dados do levantamento de setembro de 1994, com dados corrigidos manual e automaticamente.

3.1 - Módulo 1: Amendoim das Águas, Feijão das Águas e Milho

No questionário de setembro, esses produtos podem ser classificados como culturas anuais com áreas solteiras e consorciadas. Os itens analisados para cada uma delas são: alqueires que pretende plantar em cultura solteira, quilos de semente melhorada a ser utilizada em cultura solteira, quilos de semente comum a ser utilizada em cultura solteira, quilos de semente melhorada a ser utilizada em cultura consorciada e quilos de semente comum a ser utilizada em cultura consorciada.

A rotina de correções foi iniciada selecionando-se, do arquivo original (extensão.dbf), os questionários com informações referentes a uma das culturas acima, criando-se um Sasdataset[12]. Em seguida, foram separados os dados de utilização de sementes por unidade de área que respeitavam limites históricos previamente estabelecidos[13] para servirem ao cálculo do uso médio de sementes por Divisão Regional Agrícola (DIRA) e por estrato de DIRA[14].

A primeira correção inserida neste Módulo foi a eliminação de áreas ou produções, provavelmente consorciadas, coletadas como solteiras. No caso de áreas solteiras e uso de sementes em consórcio, esse erro foi caracterizado como tipo C (pergunta não compreendida), já que somente áreas solteiras devem ser informadas[15]. Se respondida apenas quanto à quantidade de sementes utilizadas em cultura solteira, o erro envolvido foi provavelmente do tipo E (erro de linha). A correção proposta para o erro tipo C foi a exclusão da área, e para o tipo E, a exclusão de semente solteira e inclusão como consorciada, segundo a especificação inicial (comum ou melhorada) informada. Em seguida, o programa calculou novamente o uso de sementes por área.

A segunda correção proposta visou eliminar o erro tipo A (resposta incompleta) através de comparações com informações anteriores do mesmo imóvel rural e referentes ao ano agrícola em questão. Para os produtos do Módulo 1, o questionário levado a campo em junho já inclui as áreas que seriam plantadas. Portanto, havendo dados de sementes solteiras (além de consorciadas), sem a correspondente área plantada, a rotina verificou se a mesma existia no arquivo de junho[16] e, se isso ocorresse, efetuava-se novamente o teste de uso de sementes por área. Caso os valores de área anterior e de uso de sementes atuais fossem consistentes, o registro de área atual seria igualado ao anterior. Conforme PINO (1986), o método de consulta às informações anteriores garante a sobrevivência de grupos. Em outros meses, é possível ampliar o número de perguntas passíveis de serem preenchidas com dados anteriores, para os produtos do Módulo 1.

A terceira correção iniciou-se recalculando o uso de sementes por área e objetivou a exclusão de erros tipo A a partir dos valores médios por DIRA e por estrato de DIRA, obtidos antes de qualquer alteração nos valores originais. Nos questionários em que ainda havia áreas sem as sementes correspondentes e vice-versa, os campos complementares foram preenchidos com base nas relações médias dos imóveis do mesmo estrato e mesma DIRA, ou, caso esse dado não existisse, com base nos valores médios dos imóveis da DIRA. Assim, no caso de área sem se-

mentes, essa quantidade foi estimada multiplicando-se a área pelo uso médio de sementes no estrato/DIRA; como era desconhecido o tipo de semente (comum ou melhorada), esse dado foi inserido no arquivo através de um novo código representativo do uso de sementes sem qualquer discriminação. Quando foi necessário estimar áreas, as quantidades de sementes foram, naturalmente, divididas pelas médias.

Um dos erros mais comuns nesses levantamentos é a informação da área como número inteiro, uma vez que o questionário deve ser preenchido com duas casas decimais (erro tipo C, pergunta não compreendida). Uma conferência inicial é efetuada antes da digitação dos dados, tentando-se eliminar esse problema, entretanto, é possível que aconteçam algumas falhas. Por isso, a quarta correção inserida no módulo separou questionários com áreas inferiores a 0,2 alqueire, cuja quantidade de sementes informada por área fosse superior ao limite máximo permitido. Em seguida, simulou-se uma correção multiplicando a área original por 100 (o que é equivalente a acrescentar duas decimais), e o teste de uso de sementes foi novamente efetuado. Quando a área corrigida levou a quantidade de sementes por área para o intervalo admissível, a correção foi mantida.

A quinta e última correção deste Módulo iniciou-se, como nas demais, pelo recálculo do uso de sementes por área; procurou-se detectar e corrigir erros tipo F (resposta fora dos limites). Esse tipo de erro é mais sujeito a discussões e restrições. Muitas vezes é possível a ocorrência de *outliers*, valores que apesar de fora dos padrões considerados normais foram informados corretamente. A distinção entre esses casos e erros não é, infelizmente, muito fácil. Uma maneira de diminuir a possibilidade de substituição de valores corretos é o emprego de limites históricos muito amplos para os testes. Dessa forma, embora esteja se aumentando o risco de erros não corrigidos, diminui-se o de alterar informações certas.

Essa situação de erro tipo F envolve duas formas de correção: manter as áreas e alterar o uso de sementes, ou manter a quantidade de sementes e alterar áreas. Optou-se pela primeira alternativa, já que parece mais razoável o produtor conhecer melhor a área que pretende plantar do que a semente a utilizar naquela área. Mesmo após efetuado o plantio, espera-se que o agricultor, que na maioria dos casos tem contato freqüente com a área da cultura, tenha melhores condições de responder sobre ela do que sobre as sementes empregadas em determinado instante. O procedimento de correção empregado é semelhante ao da terceira fase. Nos questionários em que o uso de sementes não obedecia aos limites impostos foi mantida a área, e a quantidade de sementes (comum ou melhorada, como originalmente informado) foi calculada multiplicando-se a área pelo uso médio de sementes do estrato/DIRA, ou da DIRA.

Após as correções, o programa relacionou imóveis em que ainda existiam áreas sem o correspondente uso de sementes e vice-versa, para que fosse efetuada correção manual[17]. A seguir foram listadas todas as informações dos questionários, para eventual inspeção visual, e criado um arquivo Sasdataset, convertido para dbf, que foi enviado para processamento final (expansão da amostra, cálculo de médias, variâncias, etc.).

Para permitir uma visão sintética do conjunto de operações desenvolvidas para correção dos produtos do Módulo 1, elaborou-se um fluxograma das mesmas (Figura 1).

3.2 - Módulo 2: Algodão, Milho Safrinha e Soja Safrinha

As culturas, cujas correções foram agrupadas neste Módulo, podem ser classificadas como lavouras anuais com áreas solteiras. A primeira delas, em fase de intenção de plantio, informa, portanto, áreas a serem cultivadas e uso provável de sementes. Já para as duas últimas culturas, em época de colheita, levantam-se áreas plantadas e quantidades colhidas. A diferença mais acentuada em relação ao Módulo 1 é a inexistência da primeira correção, já que não há possibilidade de erros tipo C e E na forma como ocorriam para as lavouras daquele Módulo. Após os cálculos das relações médias, passou-se diretamente à tentativa de preencher lacunas dos erros tipo A, através de comparações com o arquivo de junho. Para o algodão só foi

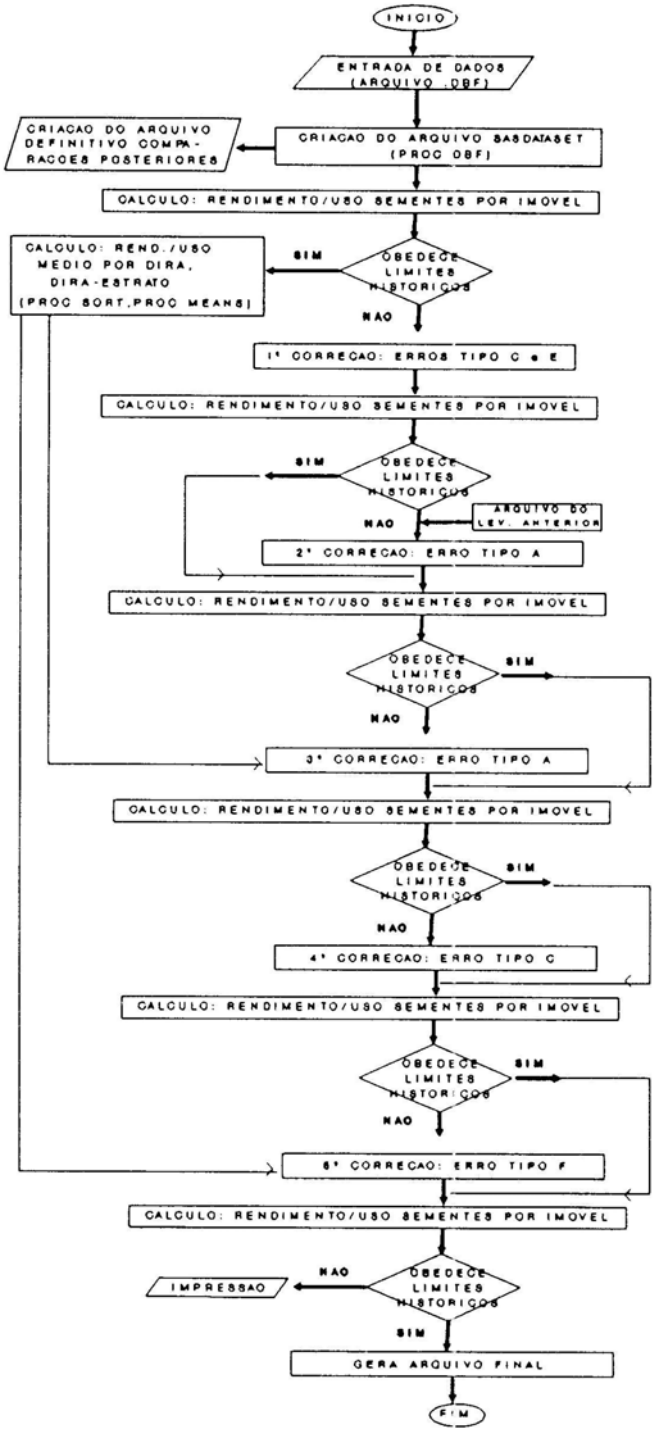

FIGURA 1 – Fluxograma das Correções do Módulo 1.

possível preencher lacunas de área, enquanto milho e soja safrinha permitiram que produções do levantamento anterior fossem aproveitadas. Devido à severa seca ocorrida no Estado de São Paulo, o limite inferior de produtividade aceita foi reduzido para apenas uma saca por alqueire. Assim, qualquer informação de produção inferior ao limite máximo foi considerada no cálculo da média para essas duas culturas. Foram, também, mantidas sem alterações áreas de até 5 alqueires sem informação de produção.

Nos demais casos, após essa primeira correção, as seguintes foram similares às terceira, quarta e quinta correções do Módulo 1.

3.3 - Módulos 3 e 4: Arroz e Feijão de Inverno

Os Módulos 3 e 4, para o levantamento de setembro, correspondem às culturas do arroz e do feijão de inverno, cujas áreas, ao contrário das lavouras dos Módulos 1 e 2, são separadas em irrigadas e não irrigadas. Foi necessário, portanto, calcular relações médias tanto para os cultivos irrigados como para os não irrigados, através de dois programas distintos, porque o número de questões, os códigos das mesmas e alguns testes diferem nas duas culturas. O limite inferior da produtividade do feijão de inverno foi reduzido para uma saca devido à seca que atingiu o Estado durante o ciclo da cultura. Por esse mesmo motivo, áreas de até 5 alqueires, sem produção, não sofreram correções.

Na primeira correção, a diferença em relação ao Módulo 1 é a ocorrência de erros tipo E, referentes a produções de lavouras de feijão irrigadas informadas como provenientes de lavouras não irrigadas e vice-versa. A correção efetuada buscou tornar a produção condizente com o tipo de área respondida.

A partir da segunda correção, foi necessário criar um código para área de arroz sem discriminar o uso de irrigação, já que a informação presente no levantamento de junho não fazia tal distinção, e as sementes a serem empregadas em cultura solteira são levantadas agregadamente em setembro. Como os demais passos das correções foram basicamente os mesmos do Módulo 1, não se considerou necessário apresentar maiores detalhes.

3.4 - Módulo 5: Soja

As questões referentes à cultura da soja, no levantamento de setembro, não são similares às das outras lavouras. As áreas levantadas são somente solteiras, e as sementes utilizadas, discriminadas em comuns e melhoradas. Por esse motivo, o roteiro de correções, embora houvesse seguido de perto o do Módulo 2, a partir de sua segunda rotina incorporou aspectos do programa do Módulo 1, com a criação de um código genérico de sementes.

3.5 - Módulos 6 e 7: Café e Laranja

As culturas perenes e semi-perenes presentes no levantamento objetivo são café, laranja e cana-de-açúcar. Em setembro, entretanto, não existem questões sobre cana, e das referentes ao café, apenas foi possível efetuar teste de consistência interna para a renda obtida no benefício. Como nesse caso é utilizada somente a média geral do Estado para transformar as sacas produzidas de café em coco em sacas de café beneficiado, foi desenvolvida uma rotina muito simples, que excluiu valores acima e abaixo dos limites históricos (16 a 23kg beneficiados por saca de 40,5kg de café em coco)[18].

As informações sobre laranja, por outro lado, levaram à elaboração de um programa que com alterações mínimas poderá ser empregado para café e cana em outros meses. Para essa cultura, após os cálculos das produtividades médias por pé, as corre-ções assemelharam-se às efetuadas no Módulo 1: pro-curou-se, na primeira correção, eliminar erros do tipo C e do tipo E, corrigindo pés novos informados como número total de pés e vice-versa, produção na linha referente a pés novos, número total de pés como pés novos e informações de pés novos sem existir o número total de pés. As demais correções também seguiram a lógica envolvida nas do Módulo 1, exceto quanto à quarta correção, que não foi efetuada neste módulo devido ao número de pés ser informado sem casas decimais.

3.6 - Módulo 8: Produção de Leite

Este módulo apresenta características distintas dos anteriores, merecendo maiores detalhes em sua descrição. As questões sobre leite, nos questioná-

rios do levantamento objetivo, compreendem vários itens: número de vacas ordenhadas ontem, litros de leite produzidos ontem, número médio de vacas ordenhadas por dia, produções mensais no mês do levantamento e nos dois ou três meses anteriores. Portanto, estão disponíveis mais informações para servirem de base a eventuais correções. Por outro lado, não há justificativa mais sólida para efetuar comparações com o levantamento anterior, já que o número médio de vacas ordenhadas pode ser alterado periodicamente.

A primeira rotina de correção, neste módulo, foi a eliminação de erros tipo A a partir do número de vacas ordenhadas e produções de outros meses. Priorizaram-se os meses mais próximos ao item com preenchimento incompleto. Então, se isso ocorresse nas informações de produção diária, a rotina buscaria preencher as lacunas de acordo com a seguinte escala de prioridades: dados de setembro, ou de agosto, ou de julho, ou, por último, de junho. Para o item produção de setembro, a ordem de prioridades de correção foi: dados diários, ou de agosto, ou de julho, ou, finalmente, de junho. Nos demais meses, procurou-se completar as respostas com valores do mês imediatamente posterior, ou imediatamente anterior, ou de dois meses a frente, ou de dois meses atrás; por último, tomou-se como base dados de produção diária.

Na segunda correção, após o recálculo das produtividades médias diárias e mensais, os casos ainda existentes de erros tipo A foram completados a partir dessas relações médias, conforme descrito no item 3.1.

Como terceiro passo, após o novo cálculo das produtividades médias, procurou-se corrigir uma série de erros mais comuns, como informar a produção média diária nos campos em que se solicita a produção total do mês (erro tipo C). Nesse caso, quando alguma produtividade média mensal esteve abaixo do limite inferior, a produção foi multiplicada pelo número de dias do mês, que é também o procedimento empregado na correção manual.

Na última correção procurou-se eliminar erros tipo F, com os novos valores de produções sendo estimados a partir das produtividades médias por estrato/DIRA ou por DIRA, conforme descrito com detalhes na quarta correção do Módulo 1. Assim como nos outros módulos, no final da rotina é emitida listagem de eventuais situações remanescentes de resposta incompleta, em seguida são listados todos os imóveis do arquivo, e, por último, é criado um arquivo final Sas-dataset, convertido para dbf, para execução dos programas de expansão da amostra.

3.7 - Comparações entre os Resultados das Depurações Automática e Manual

Comparando-se os resultados expandidos para a maioria dos itens do levantamento de setembro, percebe-se que na quase totalidade dos casos as duas correções conduziram a resultados que não diferem estatisticamente, ou seja, os intervalos de confiança dos dois dados se sobrepõem[19] (Tabela 1).

A primeira diferença essencial é a quantidade de sementes, sem discriminação, empregadas em cultura solteira para o amendoim das águas[20]. Esse código de sementes não aparece no questionário, sendo criado durante o processo de correções para informações de área sem o respectivo uso de sementes. Parte dessa diferença deve-se ao método empregado: enquanto a correção manual usa a média geral da DIRA como base de cálculo, na automática, a base é a média de estrato da DIRA correspondente, cujas vantagens foram anteriormente descritas.

Para o amendoim das águas, as sementes melhorada e comum, utilizadas em cultura solteira, tiveram os maiores valores provenientes da correção manual, sendo possível que os técnicos que as efetuaram tenham optado, em alguns casos, por discriminar o tipo de semente ao invés de inseri-las no código mais genérico. Os valores médios, encontrados para o uso de semente, foram em torno de 313kg/alq. nos dados da correção manual e cerca de 325kg/alq. nos da automática.

Para a cultura do arroz, a quantidade de semente consorciada é maior no dado expandido oriundo da correção automática. Além da diferença de base, antes referida, isso deve-se também ao critério seguido na correção automática e descrito anteriormente de eliminar áreas solteiras se a semente fosse informada no campo de cultura consorciada. Na correção manual, o técnico não tem conhecimento desse fato a partir das listagens de erros, aparecendo apenas uma situação de erro de numerador nulo, ou

TABELA 1 - Comparações entre Dados Expandidos Provenientes das Correções Manual e Automática, Estado de São Paulo, Setembro de 1994

(continua)

Produto	Variável	Unidade	Correção manual (A)				Correção automática (B)				(B)/(A)	t de Student[1]
			Valor	Int. de confiança		Erro amost. (%)	Valor	Int. de confiança		Erro amost. (%)		
				Lim. inf.	Lim. sup.			Lim. inf.	Lim. sup.			
Algodão em caroço	Área	mil alq.	72,51	62,22	82,80	14,19	72,53	62,24	82,83	14,19	1,00	0,00
	Quant. semente	mil sacas	210,26	179,88	240,65	14,45	209,25	178,91	239,59	14,50	1,00	-0,03
Amendoim das águas	Área - cult. solteira	mil alq.	9,65	7,63	11,67	20,94	9,50	7,49	11,52	21,19	0,98	-0,07
	Quant. semente comum - cult. solteira	milh. kg	1,10	0,59	1,60	45,99	1,03	0,55	1,51	46,49	0,94	-0,14
	Quant. semente melhorada - cult. solteira	milh. kg	1,89	1,48	2,30	21,53	1,83	1,42	2,24	22,23	0,97	-0,15
	Quant. semente melhorada - cult. consorc.	mil kg	14,70	...	...	...	14,70	...	...	...	1,00	-
	Quant. semente - cult. solteira	mil kg	36,82	...	...	...	235,64	...	...	...	6,40	-
Arroz em casca	Área irrig. - cult. solteira	mil alq.	4,31	3,80	4,83	12,00	4,33	3,82	4,85	11,94	1,00	0,04
	Área não irrig. - cult. solteira	mil alq.	29,30	25,01	33,58	14,62	28,64	24,40	32,89	14,81	0,98	-0,15
	Quant. semente comum - cult. solteira	milh. kg	0,69	0,62	0,77	10,69	0,70	0,63	0,78	10,80	1,01	0,10
	Quant. semente melhorada - cult. solteira	milh. kg	1,56	1,40	1,73	10,54	1,50	1,34	1,66	10,55	0,96	-0,39
	Quant. semente comum - cult. consorc.	mil kg	91,23	...	...	...	105,01	...	...	...	1,15	-
	Quant. semente melhorada - cult. consorc.	mil kg	149,25	111,24	187,27	25,47	210,85	172,83	248,86	18,03	1,41	1,62*
	Quant. semente - cult. solteira	mil kg	316,59	311,33	321,85	1,66	539,62	529,05	550,20	1,96	1,70	21,09*

[1]Os asteriscos assinalam valores maiores do que 1 (em módulo), considerados significativos.

Fonte: Instituto de Economia Agrícola (IEA).

TABELA 1 - Comparações entre Dados Expandidos Provenientes das Correções Manual e Automática, Estado de São Paulo, Setembro de 1994

(continua)

Produto	Variável	Unidade	Correção manual (A)				Correção automática (B)				(B)/(A)	t de Student
			Valor	Int. de confiança		Erro amost. (%)	Valor	Int. de confiança		Erro amost. (%)		
				Lim. inf.	Lim. sup.			Lim. inf.	Lim. sup.			
Feijão das águas	Área - cult. solteira	mil alq.	52,85	46,21	59,50	12,57	51,30	44,51	58,09	13,23	0,97	-0,23
	Quant. semente comum - cult. solteira	milh. kg	1,51	1,14	1,89	24,77	1,32	0,95	1,70	28,52	0,87	-0,50
	Quant. semente melhorada - cult. solteira	milh. kg	4,05	3,35	4,74	17,22	4,14	3,42	4,85	17,29	1,02	0,12
	Quant. semente comum - cult. consorc.	mil kg	269,46	78,82	460,11	70,75	300,27	109,63	490,91	63,49	1,11	0,16
	Quant. semente melhorada - cult. consorc.	milh. kg	1,24	0,13	2,34	89,26	1,24	0,13	2,34	89,13	1,00	0,00
	Quant. semente - cult. solteira	mil kg	237,77	...	...	...	154,96	...	...	...	0,65	-
Milho em grão	Área - cult. solteira	mil alq.	373,52	347,98	399,07	6,84	373,45	347,83	399,07	6,86	1,00	0,00
	Quant. semente comum - cult. solteira	milh. kg	1,15	0,99	1,32	14,29	1,11	0,94	1,27	14,64	0,96	-0,28
	Quant. semente melhorada - cult. solteira	milh. kg	17,25	15,80	18,70	8,41	17,43	15,97	18,90	8,40	1,01	0,12
	Quant. semente comum - cult. consorc.	mil kg	47,77	45,49	50,05	4,77	47,77	45,49	50,05	4,77	1,00	0,00
	Quant. semente melhorada - cult. consorc.	mil kg	372,12	301,90	442,34	18,87	372,13	301,91	442,35	18,87	1,00	0,00
	Quant. semente - cult. solteira	mil kg	475,46	400,15	550,77	15,84	523,64	440,23	607,06	15,93	1,10	0,58
Soja	Área - cult. solteira	mil alq.	180,33	162,77	197,90	9,74	171,19	155,76	186,61	9,01	0,95	-0,59
	Quant. semente comum - cult. solteira	milh. kg	3,23	2,26	4,19	29,95	2,92	1,95	3,88	33,11	0,90	-0,32
	Quant. semente melhorada - cult. solteira	milh. kg	34,86	31,49	38,23	9,66	33,54	30,29	36,78	9,68	0,96	-0,41

Fonte: Instituto de Economia Agrícola (IEA).

TABELA 1 - Comparações entre Dados Expandidos Provenientes das Correções Manual e Automática, Estado de São Paulo, Setembro de 1994

(continua)

Produto	Variável	Unidade	Correção manual (A)				Correção automática (B)				B/A	t de Student[1]
			Valor	Int. de confiança		Erro amost. (%)	Valor	Int. de confiança		Erro amost. (%)		
				Lim. inf.	Lim. sup.			Lim. inf.	Lim. sup.			
Feijão de inverno	Área irrig. - cult. solteira	mil alq.	150,80	35,71	265,89	76,32	14,76	13,01	16,52	11,88	0,10	-77,56*
	Área não irrig. - cult. solteira	mil alq.	35,29	28,97	41,62	17,92	28,96	25,16	32,77	13,14	0,82	-1,66*
	Produção - cultura consorciada	mil sc.60kg	6,73	...	...	...	7,09	...	...	...	1,05	-
	Produção irrigada	mil sc.60kg	...	...	...	...	775,86	668,40	883,32	13,85	-	-
	Produção não irrigada	mil sc.60kg	869,54	686,59	1.052,49	21,04	569,45	492,74	646,15	13,47	0,65	-3,91*
Milho safrinha	Área - cult. solteira	mil alq.	147,87	132,94	162,81	10,10	159,86	145,44	174,28	9,02	1,08	0,83
	Produção	milh. sc.60kg	8,35	7,33	9,36	12,21	11,42	10,28	12,56	9,98	1,37	2,70*
Soja safrinha	Área - cult. solteira	mil alq.	14,32	11,07	17,57	22,71	14,06	10,81	17,30	23,10	0,98	-0,08
	Produção	milh. sc.60kg	1,00	0,73	1,27	26,84	1,07	0,80	1,34	24,91	1,07	0,27
Café	Rendimento médio no beneficio (coco seco)	kg/sc.40kg	20,15	18,68	21,62	7,28	20,43	18,90	21,96	7,49	1,01	0,18
Laranja	N1 pés novos + adultos	milhão	270,69	234,36	307,02	13,42	268,82	232,48	305,17	13,52	0,99	-0,05
	N1 pés novos sem produção	milhão	47,70	40,81	54,58	14,43	49,16	42,07	56,24	14,42	1,03	0,21
	Produção	milh. cx.40,8kg	382,77	349,16	416,38	8,78	374,82	341,73	407,92	8,83	0,98	-0,24

[1]Os asteriscos assinalam valores maiores do que 1 (em módulo), considerados significativos.

Fonte: Instituto de Economia Agrícola (IEA).

TABELA 1 - Comparações entre Dados Expandidos Provenientes das Correções Manual e Automática, Estado de São Paulo, Setembro de 1994

(conclusão)

Produto	Variável	Unidade	Correção manual (A)				Correção automática (B)				(B)/(A)	t de Student[1]
			Valor	Int. de confiança		Erro amost. (%)	Valor	Int. de confiança		Erro amost. (%)		
				Lim. inf.	Lim. sup.			Lim. inf.	Lim. sup.			
Leite	N1 vacas ordenhadas ontem	milh. cab.	1,45	1,23	1,68	15,61	1,26	1,20	1,32	4,84	0,86	-3,23*
	N1 médio vacas ordenhadas por dia (jun.)	milh. cab.	1,33	1,27	1,40	4,86	1,34	1,27	1,40	4,85	1,00	0,09
	N1 médio vacas ordenhadas por dia (jul.)	milh. cab.	1,32	1,26	1,39	4,92	1,32	1,26	1,39	4,90	1,00	-0,02
	N1 médio vacas ordenhadas por dia (ago.)	milh. cab.	1,34	1,27	1,40	5,17	1,34	1,27	1,41	5,16	1,00	0,08
	N1 médio vacas ordenhadas por dia (set.)	milh. cab.	1,32	1,25	1,38	5,01	1,32	1,26	1,39	4,99	1,00	0,09
	Produção leite ontem	milh. litro	6,66	5,73	7,60	14,01	5,98	5,43	6,53	9,14	0,90	-1,25*
	Produção total leite (jun.)	milh. litro	203,00	186,31	219,69	8,22	203,45	186,76	220,13	8,20	1,00	0,03
	Produção total leite (jul.)	milh. litro	198,25	181,28	215,22	8,56	198,61	181,63	215,59	8,55	1,00	0,02
	Produção total leite (ago.)	milh. litro	197,29	178,07	216,50	9,74	196,80	177,61	215,98	9,75	1,00	-0,03

[1]Os asteriscos assinalam valores maiores do que 1 (em módulo), considerados significativos.
Fonte: Instituto de Economia Agrícola (IEA).

área sem informação de semente. As quantidades de sementes, sem discriminação, empregadas em cultura solteira também apresentaram resultados distintos nessa cultura e no feijão das águas[21], a exemplo do ocorrido com o amendoim das águas e provavelmente pelas mesmas razões.

As áreas expandidas de feijão de inverno diferiram enormemente. Nesse caso, certamente por erros não detectados na correção manual. Os dados publicados pelo IEA na previsão de safras de setembro de 1994 foram: área de 95,1 mil hectares e produção de 1.395 mil sacas. A correção automática fornece área[22] de 105,8 mil hectares e produção de 1.345 mil sacas. As áreas, portanto, são muito mais consistentes com o dado publicado (proveniente do levantamento subjetivo) do que a estimada a partir da correção manual. Embora a produção irrigada expandida por este último método não tenha sido processada, o que prejudica comparações, percebe-se que o dado de produção da correção automática é praticamente idêntico ao publicado.

Para o milho safrinha, a área publicada (379 mil hectares) é praticamente igual à da correção automática, mas a produção publicada (9.300 mil sacas) é 18,5% menor do que a proveniente dessa correção e 11,5% maior do que a derivada da manual.

Para o leite, as diferenças significativas entre as duas correções ocorreram no número de vacas ordenhadas e na produção de leite ontem. Comparando-se esses dados aos das previsões do número médio de vacas ordenhadas em setembro, a partir das mesmas fontes, conclui-se que o dado corrigido manualmente deve estar superestimado em cerca de 135 mil vacas, e o da automática subestimado em cerca de 67 mil. Isto indica ser possível um aperfeiçoamento na rotina inicialmente proposta para correção do leite, inserindo-se vacas e produção no dia de ontem, ausentes em questionários que tenham informado produção mensal.

4 - CONCLUSÕES E CONSIDERAÇÕES FINAIS

Foi possível desenvolver programas computacionais para correção dos itens referentes à previsão de safras, nos questionários do levantamento objetivo IEA/CATI. A partir dessa automatização proposta deve-se agilizar um dos gargalos do processo de elaboração dessas atividades, que é a fase de depuração de dados, diminuir a carga de rotinas sobre os pesquisadores envolvidos nessa atividade e, em parte, contornar o problema de rotatividade dos auxiliares de pesquisa. Também desaparecem os erros de cálculo que certamente ocorrem nas milhares de operações matemáticas efetuadas a cada levantamento, assim como amplia-se o uso de dados dos próprios levantamentos no processo de depuração, através do emprego de informações de questionários de meses anteriores.

Através de comparações entre os dados expandidos, oriundas das correções automática e manual, concluiu-se que os mesmos não diferem na grande maioria dos casos, e que as diferenças ocorridas podem, ao menos em parte, serem atribuídas a estimativas mais adequadas empregadas pelo processo automático, aliadas a uma melhor amplitude de detecção dos erros e facilidade de incorporação, pelo sistema, de soluções para as prováveis falhas. Por outro lado, um aperfeiçoamento foi proposto para versões posteriores do programa de correção para o leite, com base nas comparações de resultados.

NOTAS

[1]Parte integrante do projeto SPTC 16-021/89, apresentado no 41 Congresso Brasileiro de Usuários SAS, realizado de 4 a 7 de abril de 1995, em Piracicaba (SP). Recebido em 21/02/95. Liberado para publicação em 17/03/95.

[2]Engenheiro Agrônomo, MS, Pesquisador Científico do Instituto de Economia Agrícola.

[3]Engenheiro Agrônomo, Pesquisador Científico do Instituto de Economia Agrícola.

[4]Uma descrição dos levantamentos pode ser vista em CAMARGO FILHO et al. (1990). A metodologia empregada no dimensionamento das amostras encontra-se em CAMPOS & PIVA (1974) e em CAMARGO (1988).

[5]Sobre as vantagens e desvantagens dos levantamentos por amostragem, ver COCHRAN (1953).

[6]Ver o estudo de NEGRI NETO et al. (1988).

[7]Algodão, arroz irrigado e não irrigado, amendoim das águas e da seca, cana-de-açúcar, café, feijão das águas e da seca, de inverno irrigado e não irrigado, laranja, milho e milho safrinha, soja e soja safrinha.

[8]PINO (1989) concluiu que, em casos de falta de respostas, esse é o único procedimento que não causa problemas.

[9]Outro problema presente no levantamento objetivo, o da falta de respostas, é tratado segundo essa mesma pressuposição. PINO & CASER (1984a) estudaram o caso e apresentam métodos alternativos de lidar-se com o mesmo.

[10]Caso a opção escolhida fosse simplesmente substituir as produções faltantes pelas respectivas médias dos estratos, esse procedimento de correção levaria à subestimação das variâncias, conforme demonstrado por PINO (1986, 1989). Nesse caso, os erros de amostragem seriam artificialmente reduzidos e as estimativas obtidas, menos precisas do que aparentariam. No exemplo apresentado, entretanto, em que produções são calculadas a partir de produtos das produtividades médias pelas áreas dos questionários, não é possível saber *a priori* se as variâncias serão subestimadas ou superestimadas.

[11]Isso equivale, aproximadamente, a um nível de significância de 30%. Esse nível foi escolhido devido aos diversos erros de amostragem elevados que ocorreram no levantamento. Caso o teste empregado fosse mais rigoroso, praticamente não seriam encontradas diferenças entre os dois métodos de correção.

[12]Arquivo, em linguagem SAS, que servirá de base também para correções no levantamento posterior (no caso, em novembro), não podendo ser, por esse motivo, arquivo temporário.

[13]Para o amendoim das águas, 100 a 350 quilos por alqueire; para o feijão das águas, 40 a 300; e para o milho, 20 a 100 quilos de semente por alqueire.

[14]A segunda média é superior, tanto do ponto de vista técnico-agronômico (imóveis mais homogêneos com relação a região e tamanho), quanto estatístico, já que a amostra não é equiprobabilística. A primeira média agrupa imóveis de todos os tamanhos, embora pertençam à mesma DIRA, e deveria ser ponderada pelos fatores de expansão da amostra para ser não viesada. Infelizmente, esse dado não está disponível nessa altura do processamento. Entretanto, como a segunda média é prioritária para o processo de correção, acredita-se que o viés introduzido não deva ser elevado.

[15]Existe a possibilidade, menos provável, do erro ser relativo às sementes, quando utilizadas em culturas solteiras, porém informadas como consorciadas. Entretanto, o total de sementes consorciadas, quando da expansão da amostra para a elaboração das previsões de safras, é transformado em área solteira equivalente, através do uso médio de sementes em culturas solteiras. Portanto, a correção proposta não leva a maiores prejuízos.

[16]Optou-se por utilizar o arquivo original do levantamento anterior, isto é, antes de ser corrigido. Essa decisão visou evitar que correções efetuadas em determinado instante sejam perpetuadas.

[17]Existe a possibilidade lógica de tal ocorrência, caso todos os questionários de determinada DIRA encontrem-se fora dos limites pré-estabelecidos para os testes.

[18]Esse procedimento evita o problema de subestimação de variâncias, anteriormente citado. Conforme PINO (1989), a eliminação de não-respondentes - considerando-se que suas informações seriam, em média, semelhantes às dos respondentes - leva a estimativas de médias não viesadas, mas possivelmente menos precisas, pela diminuição do tamanho da amostra.

[19]Os limites superiores e inferiores dos intervalos de confiança foram obtidos, respectivamente, somando-se e subtraindo-se um desvio-

padrão às estimativas obtidas; significam que a probabilidade do valor verdadeiro ficar entre esses dois limites é de, aproximadamente, 68% (STEVENS, 1951). Verificar se os intervalos dos dois métodos de correção sobrepõem-se é equivalente a efetuar o teste *t* anteriormente descrito.

[20]Nesse caso, apesar da impossibilidade de elaboração do teste correspondente - o programa de expansão da amostra não estima as variâncias para os casos de respostas únicas nos estratos - a diferença proporcional entre as duas estimativas foi a maior ocorrida no levantamento de setembro.

[21]Nesse último caso, vale a observação efetuada na nota anterior, referente ao amendoim das águas.

[22]Os dados de área presentes nos questionários da objetiva são coletados e processados em alqueires (paulistas), apresentados em mil alqueires; para transformá-los em hectares, é necessário multiplicar por 2,42 (Tabela 1).

LITERATURA CITADA

CAMARGO, Milton N. **Amostra para previsão e estimativa das safras agrícolas do Estado de São Paulo em vigor a partir de junho de 1981.** São Paulo, IEA, 1988. 75p. (Relatório de Pesquisa, 27/88)

CAMARGO FILHO, Waldemar P. et al. **Estatísti-cas de produção agrícola no estado de São Paulo.** São Paulo, IEA, 1990. 1.v.

CAMPOS, Humberto & PIVA, Luiz H. O. Dimensionamento de amostra para estimativa e previsão de safra no Estado de São Paulo. **Agricultura em São Paulo,** SP, **21**(3):65-88, 1974.

COCHRAN, William G. **Sampling techniques.** New York, J. Wiley, 1953. 330p.

GOMES, Frederico P. **Curso de estatística experimental.** 4.ed. Piracicaba, ESALQ/USP, 1970. 430 p., tab.

GRILICHES, Zui. Economic data issues. **In:** ___ & INRILIGATOR, Michael D. **Handbook of econometrics.** Amsterdam, North-Holland, 1986. p.1465-1514.

NEGRI NETO, Afonso et al. Custo e benefício social de previsões e estimativas de produção agrícola: o valor da informação. **Agricultura em São Paulo, 35**(1):37-49, 1988.

PINO, Francisco A. Análise do viés em algunsprocedimentos para falta de resposta e para erros de resposta em levantamentos por amostragem. _____, SP, **36**(2):147-153, 1989.

_____. Detecção e correção de erros em levantamentos agrícolas. **Pesquisa Agropecuária Brasileira,** Brasília, **21**(9): 975-985, set. 1986.

_____. & CASER, Denise V. **Análise de erros não amostrais em levantamentos para previsão e estimativas de safras do Estado de São Paulo.** São Paulo, IEA, 1984a. 25p. (Relatório de Pesquisa, 10/84).

_____ & _____. **Falta de resposta em levantamentos por amostragem:** um estudo de caso. São Paulo, IEA, 1984b. 25p. (Relatório de Pesquisa, 08/84)

_____. & OSSIO, Jimenez H. G. **Um método para depuração de erros não amostrais em dados obtidos por levantamento de campo.** São Paulo, IEA, 1975. (não publicado).

STEVENS, Wilfred L. **Estimativa e previsão de safras através de um levantamento por amostragem.** São Paulo, Secretaria da Agricultura/PDV, 1951. 13p.