ESTIMATIVA DE SAFRA DE LARANJA EM 2008: um suco amargo[1]

Francisco Alberto Pino[2]
Vera Lúcia Ferraz dos Santos Francisco[3]

1 - INTRODUÇÃO

Ao longo de décadas a citricultura tem sido uma das principais atividades do agronegócio paulista, bem como uma das principais atividades exportadoras do país. Durante esse tempo, o Instituto de Economia Agrícola (IEA) tem provido o setor com estatísticas sistemáticas de produção. O auge dessas estatísticas foi atingido nos anos 1970, chegando a três fontes distintas de dados: o levantamento geral por amostragem, incluindo todos os produtos e conhecido por levantamento objetivo (CAMPOS; PIVA, 1974); o levantamento por amostragem específico para a citricultura[4]; e o levantamento em que engenheiros agrônomos regionais informam sobre número de pés e produção num município inteiro, conhecido por levantamento subjetivo (PINO, 2001a). Entretanto, com o passar do tempo, esse serviço deteriorou-se: o levantamento específico de citricultura foi o primeiro a ser interrompido, depois foi o levantamento objetivo (com breves períodos de retorno), enquanto o levantamento subjetivo foi o único que perdurou, sendo na primeira década do século XXI a base dos dados de previsão e estimativa de safras do IEA.

Desde sua criação, a Secretaria de Agricultura e Abastecimento do Estado de São Paulo (SAA) tem produzido estatísticas agrícolas, sendo marco importante a realização do primeiro censo agropecuário, em 1905 (PINO, 2005). Ao longo de um século, as estatísticas agrícolas estaduais têm passado por altos e baixos, por momentos de excelência técnico-científica e por momentos em que deixaram de ser prioridade por parte do governo estadual. Ao final da primeira metade do século XX, o IEA herdou essa atribuição dentro da SAA, passando a publicar periodicamente previsões e estimativas de safras estaduais para os principais produtos. Entretanto, desde os primórdios desse serviço, houve polarização entre os defensores do sistema por amostragem, seguidores de Salomão Schattan (IEA, 2003; SCHATTAN, 2003; PINO, 2004), e os defensores do sistema subjetivo, seguidores de Mario Zaroni (IEA, 1978). Esse enfoque diferencial, somado a fatores corporativos, como a transformação do trabalho técnico-científico em atividade de rotina burocrática e a preferência pela forma mais fácil de executar o serviço, bem como alguns casos de má administração, também contribuíram para a oscilação de qualidade dos resultados. Além disso, alterações no modelo de gestão pública, ainda que democraticamente escolhido, contribuíram para desmotivação funcional em relação às inovações e ao rigor científico, causando problemas na elaboração de estatísticas ao longo de duas décadas. Este último processo também ocorreu na esfera federal, que se traduziu no adiamento dos censos agropecuários (PINO, 2006) e na extinção da área técnica do Instituto Brasileiro do Café (IBC), que respondia por estatísticas de boa qualidade para o setor cafeeiro, sistema do qual o IEA fazia parte. A falta de números oficiais para a cafeicultura levou à profusão de conjecturas e tentativas de adivinhação que podem ter causado prejuízos ao setor e ao país, tendo havido até propostas para a compatibilização das diferentes conjecturas (MORICOCHI; PINO; VEGRO, 1995). Posteriormente, as estatísticas cafeeiras passaram a ser feitas por amostragem probabilística, pelo menos nos Estados de São Paulo (PINO; FRANCISCO; LORENA NETO, 2001) e Paraná, sendo o resultado nacional coordenado nos primeiros anos pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e mais recentemente pela Companhia Nacional de Abastecimento (CONAB), quando o trabalho foi estendido a nível nacional (FRANCISCO et al., 2010).

[1]Os autores agradecem a colaboração de Priscilla Rocha Silva Fagundes, Pesquisadora Científica do IEA, na fase de levantamento de dados. Registrado no CCTC, IE-51/2011.

[2]Engenheiro Agrônomo, Doutor, Pesquisador Científico do Instituto de Economia Agrícola (e-mail: pino@iea.sp.gov.br).

[3]Estatística, Pesquisadora Científica do Instituto de Economia Agrícola (e-mail: veralfrancisco@iea.sp.gov.br).

[4]Esse levantamento não se encontra descrito na literatura, mas sua metodologia é análoga à utilizada no levantamento por amostragem específico para avicultura, descrito em Piva et al. (1975).

No caso da citricultura, peculiaridades do setor podem também ter contribuído para o processo de deterioração das estatísticas. No nível produtivo, trata-se de um oligopsônio, com muitos produtores (da ordem de 20 mil ao longo do tempo[5]) e poucos compradores de matéria-prima (menos de 20 indústrias desde 1963, atualmente menos de 10), sendo que alguns destes últimos possuem ramificações nas duas principais regiões produtoras de laranja para fabricação de suco concentrado do mundo, o Estado de São Paulo (Brasil) e o Estado da Flórida (EUA). Com as exceções usuais daqueles que usam esse tipo de informação na tomada de decisões, muitos produtores rurais de citros não valorizam as estatísticas agrícolas, embora algumas lideranças do setor, eventualmente, acreditem que a divulgação de dados subestimados de produção agrícola possa elevar o preço recebido pelos produtores ou, ao contrário, que números altos demais possam "provocar um efeito negativo da previsão sobre o mercado, mesmo considerando os baixos estoques" (ASSOCITRUS, 2011). Por outro lado, as indústrias mais modernas têm seus próprios sistemas de estatísticas, o que lhes permite fazer previsões e estimativas de safra com precisão supostamente razoável ou boa, uma vez que tais informações são de uso empresarial exclusivo e não públicas. Isso lhes permite aguardar a publicação de dados oficiais para então assumir uma de duas posições, de acordo com seus interesses. Se os dados publicados lhes forem favoráveis, permanecerão em silêncio, caso contrário, vão atacá-los vigorosamente. Esses procedimentos por parte de produtores rurais e de industriais, na verdade, são perfeitamente compatíveis com a lógica capitalista.

Com tantos fatores desfavoráveis, não é surpreendente o declínio das estatísticas agrícolas estaduais, em especial e das estatísticas citrícolas. Uma das razões pelas quais essas estatísticas ainda mantêm algum nível de confiabilidade e aceitação pública é sua transparência metodológica. De fato, ao longo do tempo, o IEA tem sempre publicado não apenas os dados levantados, mas também a descrição detalhada da metodologia utilizada, o que não acontece com eventuais estatísticas alternativas, exceto as oficiais, como é o caso daquelas obtidas pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Além disso, o IEA publica eventuais críticas e avaliações sobre seus próprios dados, bem como desenvolvimentos metodológicos.

Ao longo do tempo, alguma reação tem sido esboçada para melhorar tais serviços. Em 1995/1996 e em 2007/2008, a SAA, por meio do IEA e da Coordenadoria de Assistência Técnica Integral (CATI), realizou dois censos agropecuários, conhecidos por projeto LUPA ou censo agropecuário paulista (PINO et al., 1997; PINO, 2000; TORRES et al., 2009). Deles deveria resultar um sistema de estatísticas, que chegou a ser proposto pela Comissão de Estatísticas da SAA, mas feneceu por falta de apoio (em alguns casos, oposição) de alguns dirigentes (PINO; FRANCISCO, 2001; PINO, 2000). Em 2009, produziu-se no IEA importante trabalho para determinar a área de citros mediante o uso de imagens de satélite, obtendo-se o valor de 521.125,5 ± 11.716,1 ha para laranja, limão e tangerina (MARTINS, 2009). Esse trabalho deveria ser continuado em anos seguintes. A utilização de modelos matemáticos para melhorar as previsões também foi sugerida e testada em algumas ocasiões (PINO; AMARO, 1986; PINO; CÉZAR; AMARO, 1992), bem como novos levantamentos amostrais (SANTOS et al., 1987).

O objetivo geral do presente artigo é descrever o levantamento por amostragem utilizado para estimar e/ou prever variáveis como a produção, a área plantada e o número de plantas, bem como variáveis correlatas, para a cultura da laranja no Estado de São Paulo na safra agrícola 2007/2008 (safra industrial 2008/2009). O objetivo específico é apresentar o estado da arte do tratamento estatístico para falta de resposta.

2 - MATERIAL E MÉTODO

A população alvo foi definida como o conjunto das unidades de produção agropecuária (UPAs) de todo o Estado de São Paulo, com área positiva de laranja, constantes do censo agropecuário paulista, também conhecido por projeto LUPA (TORRES et al., 2009; FAGUNDES et al., 2010), com dados de 2007/2008.

2.1 - Amostragem

Foi utilizada uma amostra probabilística estratificada, tendo como unidade amostral a

[5]Com pessoal ocupado da ordem de 200 mil pessoas.

UPA produtora de laranja.

2.1.1 - Tamanho da amostra

O tamanho máximo da amostra foi estabelecido levando em conta os recursos humanos, materiais e financeiros disponíveis para o levantamento no campo, resultando em cerca de 500 UPAs (PINO, 2002). Por outro lado, pode-se facilmente fazer o cálculo do tamanho máximo da amostra considerando a estimativa de proporções para variáveis categóricas binárias, fixando-se o número de pontos percentuais acima e abaixo que se deseja para o intervalo de confiança da estimativa da proporção (KISH, 1965, CAMPOS; PIVA, 1974).

Seja *d* o valor desejado para a semi-amplitude do intervalo de confiança da estimativa da proporção. Neste caso, a variância desejada é dada por $V^2 = d^2 / t^2$, onde *t* é o valor da tabela da distribuição t de Student a ser utilizado. Seja S^2 a variância e *P* a proporção na população, com $S^2 = P(1-P) \leq 0{,}5^2 = 0{,}25$. Então, para *d* fixado, e *N* o tamanho da população, um limite superior para o tamanho da amostra é dado por $n = n'/(1+n'/N)$, onde $n' = S^2/V^2 \leq 0{,}25t^2/d^2$. Tomando-se valores entre 3 e 5 pontos percentuais para *d*, bem como *t* = 1,96 e *N* = 20.320[6], tais valores podem ser calculados (Tabela 1).

O tamanho de amostra pretendido, de até 500 UPAs, resulta, portanto, num desvio máximo entre 4 e 5 pontos percentuais para cima e para baixo. Pode-se afirmar que o valor real estará abaixo desses valores na maioria dos casos. Se uma amostra estratificada for utilizada, precisão ainda maior deverá ser obtida. Espera-se que a amostra funcione relativamente bem no caso de variáveis racionais, ao invés de categóricas, como é o caso de área, produção e similares[7].

2.1.2 - Esquema amostral

No delineamento amostral proposto utilizou-se estratificação das UPAs por tamanho da cultura de laranja.

Dado um critério de estratificação, a eficiência máxima ocorre quando se aumenta o número de estratos, diminuindo consequentemente o tamanho da amostra em cada um deles até o valor mínimo de dois elementos[8]. Adotou-se o tamanho da cultura (área plantada com laranja na UPA) como critério de estratificação, com dois elementos por estrato, exceto no estrato com os maiores laranjais, em que se adotou o censo, isto é, todos os elementos desse estrato devem ser levantados. As UPAs foram estratificadas de um em 1 ha até 10 ha, passando para 5 em 5 ha, e daí por diante. Após algumas simulações, o esquema amostral pretendido foi obtido.

Uma particular amostra foi sorteada pelo IEA, enquanto que assistentes agropecuários e técnicos da CATI aplicaram um questionário no campo às UPAs sorteadas, entre agosto e setembro de 2008. As estimativas foram calculadas pelas fórmulas usuais de amostragem probabilística estratificada (KISH, 1965).

2.2 - Falta de Respostas

A correção de erros de preenchimento é feita normalmente no IEA, mediante técnicas de controle de qualidade dos dados, restando somente casos de falta de resposta em que a UPA inteira deixa de responder. No presente trabalho, o teste de Little foi aplicado aos dados para identificar o mecanismo de falta de resposta (*missing values*), a qual foi tratada segundo as seguintes técnicas de imputação de dados: a) eliminação de casos não disponíveis (*pairwise delection - PD*); b) imputação condicional pela média ou imputação por regressão; c) imputações múltiplas pelo método de Monte Carlo via cadeias de Markov (MCMC); d) imputações múltiplas pelo método de regressão; e e) imputações múltiplas com maximização da esperança (*expectation maximization - EM*). A primeira é uma técnica ingênua, indireta, que ignora os dados; a segunda é uma técnica de imputação condicional; as demais são técnicas de imputação iterativa.

[6]O número de UPAs constantes em Torres et al. (2009) é de 20.720, tendo sido alterado neste trabalho para o valor apresentado. Os casos eliminados são de pomares domésticos e outros em interesse para a presente pesquisa.

[7]Ver, por exemplo, a amostra utilizada em outra cultura perene, o café (PINO; FRANCISCO; LORENA NETO, 2001).

[8]Dois elementos é o tamanho mínimo que permite o cálculo da variância amostral e, por consequência, do erro amostral, o qual permite o controle de precisão das estimativas obtidas.

TABELA 1 - Tamanho Máximo de Amostra para Estimativa de Variáveis Categóricas sobre Laranja, Estado de São Paulo, 2007/2008

Pontos percentuais	d	V^2	n' máximo	n máximo
3	0,03	0,000234	1.067	1.014
4	0,04	0,000416	600	583
5	0,05	0,000651	384	377

Fonte: Elaborada pelos autores com base em Torres et al. (2009).

3 - FALTA DE RESPOSTA: levantamento sobre o estado da arte

É oportuno fazer aqui uma revisão sobre o assunto, que se mostrou relevante no presente trabalho. Dado um processo estocástico $\mathbf{Y}$, obtém-se um conjunto de dados (*data set*) mediante um levantamento de dados estatísticos. Como se faz usualmente com informação estruturada, os dados podem ser arranjados numa matriz tridimensional

$$\mathbf{Y} = \left\{ Y_{ijt}; i = 1, 2, ..., n; \; j = 1, 2, ..., \; p; t = 1, 2, ..., T \right\}$$

onde n é o número de observações, ou registros (*records*), ou linhas num arquivo; p é o número de variáveis levantadas (ou de campos, ou de colunas num arquivo); e T é o número de instantes (anos, meses, semanas, dias, horas, etc.) no tempo. Num estudo longitudinal (*longitudinal study*); tem-se $T > 1$, enquanto em um estudo de corte transversal (*cross-sectional study*) tem-se $T = 1$, reduzindo a matriz a duas dimensões.

Diz-se que existem valores perdidos ou faltantes (*missing values*) quando não se dispõe de um ou mais dados para a matriz $\mathbf{Y}$. Por exemplo, a observação de uma dada variável num dado instante pode ter sido medida de forma errada no campo ou no laboratório, produzindo um dado faltante. Um caso particular é o de falta de resposta (*nonresponse*)[9], quando todos os dados de um registro (ou linha) da matriz estão faltando, embora neste trabalho seja utilizada a expressão falta de resposta para a designação geral do problema. Um conjunto de dados sem valores perdidos é dito completo, enquanto um com valores perdidos é dito incompleto.

Há décadas o IEA considera o problema de dados faltantes como erro no preenchimento de questionário no campo, que é tratado por meio de procedimentos de detecção (*error detection*) e correção (*error correction*) de valores errôneos (*erroneous values*) que constituem o controle de qualidade dos dados (PINO, 1986; FRANCISCO et al., 1998; BRUNI, 2004), bem como do problema de falta de resposta em levantamentos (PINO; CASER, 1984; PINO; FRANCISCO, 2001).

Neste estudo, n representa o tamanho da amostra (i.e., o número de UPAs da amostra), com p variáveis e $T = 1$. Por isso, deste ponto em diante, considera-se o caso de um estudo de corte transversal, eliminando-se, por simplicidade, a componente temporal[10].

Certamente, a melhor maneira de lidar com valores perdidos é evitar que eles ocorram, mediante cuidadosos planejamento e levantamento de dados (DESARBO et al., 1986, apud OLINSKY; CHEN; HARLOW, 2003)[11]. Entretanto, o problema aparece em todos os campos científicos, com maior ou menor gravidade e devido a um grande número de causas. Ao longo dos últimos 20 anos, tem surgido grande número de propostas e de algoritmos computacionais para tratar da questão de valores perdidos (WA-

[9]As expressões "valores perdidos" (*missing values*) e "valores faltantes" referem-se ao problema geral de não obtenção de dados, seja em levantamentos de dados mediante entrevista (amostral ou censitário), seja em medições ou observações (experimentais, laboratoriais, ou outras). A expressão "falta de resposta" (*nonresponse*) refere-se a "valores perdidos" no caso particular em que dados são obtidos em levantamentos de dados mediante entrevista.

[10]Para estudos de valores perdidos em dados longitudinais, ver Twisk e Vente (2002), Engels e Diehr (2003). O método mais simples neste caso consiste em repetir o valor anterior para o valor perdido (*last value carried forward - LVCF*), o que pode ser feito, por exemplo, com a área plantada em levantamentos de safras sucessivos (ao longo de um ano, para uma cultura anual, e até entre dois anos consecutivos, para culturas permanentes). Outros métodos são: o de interpolação linear (*linear interpolation*), que imputa a média entre o valor anterior e o valor posterior; o de regressão longitudinal individual; e o de regressão longitudinal populacional. Embora não mencionado na literatura citada, convém citar a possibilidade de usar modelos de séries temporais para prever valores perdidos no tempo.

[11]DESARBO, W.S.; GREEN, P.E.; CARROLL, J.D. Missing data in product-concept testing. **Decision Sciences**, v. 17, p. 163-185, 1986.

SITO; MIRKIN, 2006); no entanto, ainda não se dispõe de uma teoria geral (IACUS; PORRO, 2007).

3.1 - Mecanismo e Padrões da Falta de Resposta

A falta de resposta apresenta alguns mecanismos que podem ser reconhecidos, conforme proposto por Little e Rubin (1987 apud TWISK; VENTE, 2002; WASITO; MIRKIN, 2006; SENTAS; ANGELIS, 2006; SONG; SHEPPERD, 2007; QIN et al., 2009)[12] e Hair Junior et al. (2005), a saber:

a) falta de resposta completamente aleatória (*missing completely at random - MCAR*) ou dados perdidos completamente ao acaso, quando ela é independente tanto dos dados observados quanto dos dados não observados, i.e., os valores perdidos de uma variável não são relacionados aos valores de quaisquer outras variáveis, perdidas ou não. Neste mecanismo, dados de respondentes e dados de não respondentes são indistinguíveis;
b) falta de resposta aleatória (*missing at random - MAR*) ou dados perdidos ao acaso, quando ela depende dos dados observados, mas independe dos dados não observados, i.e., os valores perdidos de uma variável não dependem de valores dela mesma, mas dos valores de outras variáveis. Trata-se de um mecanismo intermediário entre os outros dois, que são casos extremos. Neste mecanismo, embora dados de respondentes e dados de não respondentes sejam diferentes, é possível predizer o padrão de perda de valores de uma variável a partir de outras variáveis do conjunto de dados;
c) falta de resposta não aleatória ou não ignorável (*missing not at random - MNAR* ou *non-ignorable missing - NIM*), quando ela depende dos dados não observados, i.e., os valores perdidos de uma variável dependem inclusive de valores dela mesma. Este tipo é comum quando os respondentes não querem revelar alguma coisa pessoal, ou impopular, ou confidencial, principalmente no caso de empresas (OLINSKY; CHEN; HARLOW, 2003), que preferem não divulgar certas informações (este último motivo é especialmente importante no caso da estimativa de safra de laranja) ou quando parte dos informantes não sabe responder. Neste mecanismo, não é possível predizer o padrão de perda de valores de uma variável a partir de outras variáveis do conjunto de dados.

Essas definições podem ser formalizadas considerando a matriz indicador

$$\mathbf{M} = \left\{ m_{ijt}; i = 1, 2, ..., n; \; j = 1, 2, ..., k; t = 1, 2, ..., T \right\}$$

onde $m_{ijt} = 1$ se Y_{ijt} é um valor perdido e $m_{ijt} = 0$, caso contrário. Considerando-se o conjunto *F* dos valores perdidos de **Y** e o conjunto *R* dos valores não perdidos de **Y**, então, o mecanismo da falta de resposta pode ser caracterizado pela distribuição condicional de **M** dado **Y**, a saber, $f\left(\mathbf{M}\middle|\mathbf{Y},\theta\right)$, onde θ é um vetor de parâmetros desconhecido (SENTAS; ANGELIS, 2006; SONG; SHEPPERD, 2007). Na falta de resposta completamente aleatória (MCAR) essa distribuição não depende de **Y**, i.e.,

$$f\left(\mathbf{M}\middle|\mathbf{Y},\theta\right) = f\left(\mathbf{M}\middle|\theta\right) \text{ para todo } \mathbf{Y}.$$

Na falta de resposta não aleatória ou não ignorável (NI), essa distribuição não é independente de **Y**, i.e.,

$$f\left(\mathbf{M}\middle|\mathbf{Y},\theta\right) \neq f\left(\mathbf{M}\middle|\theta\right) \text{ para todo } \mathbf{Y} \text{ e } f\left(\mathbf{M}\middle|\mathbf{Y},\theta\right) \text{ depende de } F.$$

Na falta de resposta aleatória (MAR), essa distribuição não depende dos valores faltantes (*F*), mas pode depender dos valores observados de outras variáveis em *R*, i.e.,

$$f\left(\mathbf{M}\middle|\mathbf{Y},\theta\right) = f\left(\mathbf{M}\middle|R,\theta\right) \text{ para todo } F.$$

Para testar se a falta de resposta é completamente aleatória (MCAR), usa-se o teste multivariado de Little, disponível em alguns *softwares*: se o teste for não significativo, pode-se considerar que os dados sejam MCAR (LITTLE, 1988).

Examinando a matriz de valores perdidos **M**, é possível reconhecer alguns padrões para a falta de resposta, como o univariado e o multivariado. No padrão univariado de falta de resposta (*univariate pattern*), somente uma variável contém valores perdidos, enquanto no padrão

[12] LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**. New York, Wiley, 1987.

multivariado de falta de resposta (*multivariate pattern*), mais de uma variável contém valores perdidos (SONG; SHEPPERD, 2007). Por sua vez, este último pode ser subdividido em:

a) padrão monótono (*monotone pattern*), se as variáveis podem ser arranjadas de tal forma que se a variável Y_{iht} tem valor perdido, então, $Y_{ijt}, j = h+1,...,k$ também têm valor perdido[13].
b) padrão arbitrário (*arbitrary pattern*), se os valores perdidos podem ocorrer em qualquer lugar da matriz, sem nenhuma estrutura especial.

Os valores perdidos ou faltantes (*missing values*) podem ser tratados de maneira direta ou indireta. Na forma direta, os métodos trabalham com os dados perdidos, podendo estimar parâmetros. A maneira indireta torna completo o conjunto de dados, seja substituindo os valores perdidos por alguma forma de imputação ou atribuição, seja deixando-os de lado na análise. As técnicas indiretas são muito usadas, podendo ser classificadas em três grupos (SONG; SHEPPERD, 2007): as que ignoram, as que toleram e as que imputam dados. Todas as técnicas têm limitações, podendo ser influenciadas por valores estranhos ou atípicos (*outliers*), como mostrado num estudo sobre robustez dos métodos de imputação de dados (BRANDEN; VERBOVEN, 2008).

3.2 - Técnicas Indiretas que Ignoram os Dados

As técnicas aqui descritas consistem simplesmente em desconsiderar os casos que contêm valores perdidos, usando somente os respondentes (*R*), o que, em alguns casos, equivale a substituir os valores perdidos por zeros (TROYANSKAYA, 2001) e produzem estimativas viesadas, a menos que os valores perdidos ocorram aleatoriamente (MCAR), com grande perda de dados. Mesmo com MCAR, são recomendadas somente quando o número de valores perdidos é pequeno (SONG; SHEPPERD, 2007). Ainda assim, são amplamente usadas devido à sua simplicidade (SENTAS; ANGELIS, 2006). Há dois casos a considerar:

a) eliminação de casos completos (*listwise delection - LD*). Neste caso, todas as variáveis de um dado caso são eliminadas. No problema atual, um caso completo corresponde a uma UPA, enquanto as variáveis são a produção, a área plantada, o número de plantas de laranja, etc.;
b) eliminação apenas de casos não disponíveis (*pairwise delection - PD*). Neste caso, a eliminação ocorre somente nas variáveis para as quais os dados foram perdidos. Não é aplicável ao caso atual, já que a UPA inteira não está disponível.

3.3 - Técnicas Indiretas que Toleram os Dados

Neste caso, usa-se um enfoque probabilístico para tratar os valores perdidos, que consiste em especificar uma probabilidade para cada um dos possíveis valores. O algoritmo CART, proposto por Breiman et al. (1984), citado por Song e Shepperd (2007), é uma das formas de fazer os cálculos necessários. No problema atual, o fato de se trabalhar com variáveis contínuas dificulta a aplicação desse tipo de técnica.

3.4 - Técnicas Indiretas de Imputação de Dados

São procedimentos estatísticos que consistem em usar os dados recebidos para estimar algum tipo de modelo, o qual, aplicado aos valores perdidos, produz estimativas para substituir os *missing values*[14]. Dessa forma, o conjunto de dados fica completo para que nele sejam aplicadas as análises desejadas. São aplicados principalmente em dados numéricos, produzindo estimativas contínuas, mas o uso de regressão logística multinomial (*multinomial logistic regression - MLR*) para variáveis categóricas foi proposto por Sentas e Angelis (2006). Se feita com cuidado, a imputação leva a estimadores consistentes e testes válidos (NIELSEN, 2001). Algumas das formas mais comuns de imputar dados são apresentadas a seguir.

3.4.1 - Técnicas de imputação pela média

Compreende métodos não iterativos que usam todos os valores respondentes (*R*) para inferir sobre os não respondentes (*F*). A

[13]Ver também Liu (1995) e Nielsen (2001).

[14]Comentários sobre o assunto, nos levantamentos do IEA, podem ser encontrados em Pino (1986).

imputação pela média pode ser incondicional ou condicional.

a) imputação incondicional pela média (*incondicional mean imputation - MI*). Consiste em substituir cada valor perdido por alguma média[15] calculada sobre os valores respondidos. Tem a desvantagem de subestimar a variância, induzindo o usuário a pensar que a estimativa é mais precisa do que ocorre na realidade. É uma técnica fácil de aplicar, mas ainda ingênua (*naïve*). Apenas por simplificação, considere-se o caso de um estudo em corte transversal, com matriz de dados dada por

$$\mathbf{Y} = \left\{ Y_{ij}; i = 1, 2, ..., n; \; j = 1, 2, ..., \; p \right\}$$

com *r* respondentes e *m* = *n* – *r* não respondentes. Seja

$$s_r = \left\{ (i, \; j) \middle| i = 1,2,...,n; \; j = 1,2,..., \; p, Y_{ij} \in R \right\}$$

o conjunto de índices dos valores respondidos e

$$s_m = \left\{ (i, \; j) \middle| i = 1,2,...,n; \; j = 1,2,..., \; p, Y_{ij} \in M \right\}$$

o conjunto de índices dos valores perdidos. Tome-se uma amostra aleatória $\left\{ Y_{ij}^* \middle| i \in s_m; \; j = 1,2,..., \; p \right\}$ de tamanho *m*, selecionados com reposição a partir do conjunto de respondentes $\left\{ Y_{ij} \middle| i \in s_r; \; j = 1,2,..., \; p \right\}$. Seja $\tilde{\tilde{Y}}_{ij}$ o valor a ser imputado, i.e., o valor que vai substituir o valor perdido, para algum par (*i, j*). Definem-se três casos de imputação incondicional pela média (MOJIRSHEIBANI, 2001):

- imputação pela média dos respondentes (*mean imputation*):

$$\tilde{\tilde{Y}}_{ij} = \bar{Y}_r$$

onde $\bar{Y}_r = r^{-1} \sum_{i \in s_r} Y_{ij}$ é a média de todos os respondentes;

- imputação aleatória (*random imputation*):

$$\tilde{\tilde{Y}}_{ij} = Y_i^*$$

em que o valor perdido é substituído por um valor não perdido selecionado aleatoriamente;

- imputação aleatória ajustada (*adjusted random imputation*):

$$\tilde{\tilde{Y}}_{ij} = \bar{Y}_r + \left(Y_i^* - \bar{Y}_m^* \right)$$

onde $\bar{Y}_m^* = m^{-1} \sum_{i \in s_m} Y_{ij}^*$ é a média da amostra de respondentes.

Mojirsheibani (2001) mostrou que imputações aleatórias ou aleatórias ajustadas produzem estimativas confiáveis da função de distribuição desconhecida.

b) imputação condicional pela média ou imputação por regressão (*regression imputation - RI*). Consiste em substituir cada valor perdido por um valor previsto calculado com base num modelo de regressão que, por sua vez, é estimado sobre os valores observados (ou não perdidos). Tende a funcionar melhor do que a imputação pela média, mas também subestima a variância. As estimativas dos parâmetros são consistentes no caso de falta de resposta aleatória (NIELSEN, 2001; BUCK, 1960).

3.4.2 - Técnicas de imputação por observações similares

Compreende métodos não iterativos que usam apenas parte dos valores respondentes (*R*) para inferir sobre os não respondentes (*F*), desde que os elementos, faltante e respondente, possam ser considerados semelhantes de alguma maneira. Pode usar somente uma observação ou um grupo de observações semelhantes:

a) imputação por observação similar (*hot-deck imputation - HDI*)[16]. Consiste em substituir o valor do elemento faltante pelo valor de um elemento respondente, em que o faltante e o respondente são considerados semelhantes de alguma maneira. Ambos pertencem, porém, ao mesmo conjunto de dados. Há diver-

[15] Em alguns casos, mediana.

[16] *Hot-deck* é um termo usado no jogo de pôquer para designar um baralho do qual saem boas mãos. Alguns autores chamam de "técnicas *hot-deck* de imputação" todas aquelas que se baseiam na substituição de valores perdidos por valores de respondentes do próprio conjunto de dados no qual existem elementos faltantes, em contraposição às "técnicas *cold-deck*", que usam valores de outras fontes ou pesquisas, inclusive de pesquisas anteriores sobre o mesmo assunto (IACUS; PORRO, 2007). Por outro lado, Wasito e Mirkin (2005) chamam de *hot-deck* a imputação com o valor vizinho mais próximo e de *cold-deck* a imputação com o valor modal.

sas maneiras de escolher qual valor é o mais similar para ser tomado. A mais simples consiste em escolher aleatoriamente um valor dentre os valores possíveis, isto é, sortear um dos elementos respondentes considerados semelhantes ao elemento faltante. Também existem algoritmos como os usados na imputação por padrão similar (*similar response pattern imputation - SRPI*), que identifica o valor mais semelhante entre os respondentes e usa-o para substituir o valor perdido, ou na imputação pelos *k* vizinhos mais semelhantes (*k nearest neighbours - KNN*), que procura os *k* vizinhos mais semelhantes entre os respondentes, e usa a média ou o valor modal desses elementos para substituir o valor perdido (SONG; SHEPPERD, 2007; TROYANSKAYA, 2001); ou, ainda, na imputação pelos *k* vizinhos mais semelhantes sequenciais (*sequential k nearest neighbours - SKNN*), em que os valores imputados são usados nos passos seguintes do algoritmo (VERBOVEN; BRANDEN; GOOS, 2007).

b) substituição pela média do grupo similar (*group mean substitution*). Neste caso, usa-se a média de um grupo de elementos que sejam relativamente homogêneos em relação à variável com valor perdido (OLINSKY; CHEN; HARLOW, 2003).

c) substituição pelos *k* vizinhos mais próximos (*k nearest neighbours - KNN*). A vizinhança pode ser determinada pela distância euclidiana ou similar (TROYANSKAYA, 2001; OBA et al., 2003).

3.4.3 - Técnicas de imputação iterativa

Compreendem métodos e algoritmos que obtêm os valores a serem imputados de forma iterativa:

a) imputações múltiplas (*multiple imputation*). Consiste em imputar um valor perdido $m > 1$ vezes, adicionando um termo de erro aleatório a cada imputação. Obtêm-se, assim, *m* conjuntos de dados completos, aplica-se a mesma análise a cada um e os resultados são combinados para produzir estimativas gerais. Se os dados perdidos tiverem padrão monótono (*monotone pattern*), o método de regressão pode ser utilizado. Caso contrário, se os dados perdidos apresentarem padrão arbitrário (*arbitrary pattern*), o método de Monte Carlo via cadeias de Markov (*Markov chain Monte Carlo* - MCMC), descrito por Schafer (1997 apud DYBOWSKI, 1998)[17], pode ser utilizado.

b) maximização da esperança (*expectation maximization - EM*). Consiste em obter estimativas de máxima verossimilhança de um modelo, de forma iterativa, por meio do algoritmo EM (*expectation-maximization algorithm*), e substituir o valor perdido pelo valor esperado (SENTAS; ANGELIS, 2006).

c) substituição sequencial pelos *k* vizinhos mais próximos (*sequential k nearest neighbours - SKNN*)[18]. Usa a técnica de substituição pelos *k* vizinhos mais próximos, mas assim que um valor é imputado, ele passa a fazer parte do conjunto de dados para a imputação seguinte. Uma técnica semelhante consiste em estimar os valores perdidos mediante a minimização de um critério de covariância (*SEQimpute*)[19].

d) Decomposição em valores singulares (*singular value decomposition - SVD*). Consiste na obtenção de valores perdidos por meio de um algoritmo iterativo (*IAimpute*) que usa uma decomposição em valores singulares (VERBOVEN; BRANDEN; GOOS, 2007; TROYANSKAYA, 2001).

e) Substituição baseada em análise de componentes principais bayesiana (*Bayesian principal component analysis - BPCA*)[20]. Consiste em três processos: regressão de componentes principais, estimaç ão bayesiana e um algoritmo iterativo do tipo EM (VERBOVEN; BRANDEN; GOOS, 2007).

3.4.4 - Técnicas de imputação usando redes neurais

Consistem na utilização de redes neurais artificiais (*artificial neural networks - ANN*) para estimar os valores perdidos (WASITO; MIRKIN, 2006; OLINSKY; CHEN; HARLOW, 2003; DYBOWSKI, 1998), como mapas auto-organizáveis (*self-organizing maps - SOM*) e perceptron multicamadas (*multi-layer perceptron - MLP*) com algoritmo de aprendizagem de retropropa-

[17]SHAFER, J. L. **Analysis of incomplete multivariate data**. London: Chapman & Hall, 1997.

[18]Proposto por Kim, Kim e Yi (2004).

[19]Proposto por Verboven, Branden e Goos (2007).

[20]Proposto por Oba et al. (2003).

gação (*back-propagation*), conforme Junninen et al. (2004). Os dados conhecidos são usados para treinar um modelo a reagir de forma semelhante.

3.5 - Técnicas Diretas de Imputação de Dados

Apresentam-se, a seguir, duas possibilidades.

3.5.1 - Técnica de imputação usando equação estrutural

O uso de modelagem por equação estrutural (*structural equation modelling approach - SEM*) é a forma mais elegante de substituir valores perdidos, no dizer de Olinsky, Chen e Harlow (2003), mas ele funciona melhor quando existem somente uns poucos padrões de falta de resposta.

3.5.2 - Técnica de imputação por estimação de máxima verossimilhança

Sob a suposição de normalidade dos erros, utiliza-se a estimação de máxima verossimilhança (*full information maximum likelihood - FIML*), que tem algumas vantagens sobre os modelos de equações estruturais (OLINSKY; CHEN; HARLOW, 2003).

4 - RESULTADOS E DISCUSSÃO

Como o principal objetivo desta pesquisa era obter uma amostra funcional, o primeiro resultado a ser apresentado é a própria amostra, que tem 485 UPAs, classificadas em 69 estratos (Tabela 2). Os estratos foram numerados em ordem decrescente de área plantada com laranja. Assim, o primeiro estrato contém UPAs com mais de 300 ha de laranja e nele se realiza levantamento censitário, i.e., todas as 349 UPAs são levantadas[21]. Nos demais estratos, somente um par de elementos é levantado. No geral, fazem parte da amostra apenas 2,4% das UPAs com laranja. Contudo, elas representam 40,2% da área plantada e 57,9% do número de plantas, indicando que a amostra concentra-se nas plantações maiores e com maior adensamento[22]. Como o adensamento constitui prática recente, conclui-se que a amostra concentra-se nas plantações recentes, que permanecerão ainda por muitos anos. É provável também, que tais UPAs sejam as de melhor padrão tecnológico e com pomares realmente comerciais. Além do mais, 349 UPAs do estrato 1 são levantadas de maneira censitária, com erro amostral nulo, o que significa que 56,3% das plantas serão levantadas censitariamente[23]. Essas características distinguem o esquema amostral proposto, mostrando seu potencial para obtenção de dados de boa qualidade.

O levantamento no campo apresentou sérios problemas de falta de resposta[24]: 121 dos 485 elementos (25% da amostra) recusaram-se a fornecer dados, os quais representam 6% do total de produtores em número, porém 19% da área total com laranja. No estrato 1 (UPAs com área de laranja acima de 300 ha), a falta de resposta foi de quase 90% (Tabela 3). As principais causas alegadas foram: a) os dados solicitados são considerados segredos empresariais, tanto de industriais, quanto de produtores agrícolas; e b) os dados não podem ser fornecidos por razões contratuais (devido a cláusulas estabelecidas em contratos de compra e venda de laranja). Na verdade, algumas das grandes indústrias não permitiram que suas fazendas, nem as fazendas de fornecedores contratados, fossem examinadas pelo pessoal de campo. Como a amostra estava calcada sobre os grandes produtores rurais, a falta de colaboração deles inviabilizou completamente o levantamento. Mesmo com o

[21]Embora o limite de 300 ha para o estrato de grandes produtores tenha sido encontrado com base numa simulação matemática, ele pode ser justificado praticamente. De acordo com o Dr. Antonio Ambrosio Amaro, pesquisador aposentado do IEA e uma das maiores autoridades em economia citrícola do país, 100 mil plantas de laranja equivalem a três *kits* de máquina, sendo acima desse valor considerado grande produtor (comunicação verbal). De acordo com o censo agropecuário paulista (TORRES et al., 2009), a densidade média de plantio é de 350 plantas/ha, resultando 285,7 ha como a área acima da qual se considera grande produtor, valor muito próximo ao adotado no presente trabalho.

[22]Nesse contexto, "concentrar-se" significa que a média (do tamanho da plantação ou da densidade de cultivo) na amostra é superior à média da população.

[23]Mais precisamente, 195.866.224 plantas dentre 348.088.464 serão levantadas censitariamente (Tabela 2).

[24]As conseqüências estatísticas da falta de resposta podem ser vistas em Pino e Caser (1984) e Pino (1989).

TABELA 2 - Esquema Amostral para Previsão e Estimativa de Safra de Laranja, Estado de São Paulo, Safra Agrícola 2007/08

(continua)

Estrato	Limites		População			Amostra			% Amostra / População		
	Inferior	Superior	Número de UPAs	Área de laranja (ha)	Número de plantas	Número de UPAs	Área de laranja (ha)	Número de plantas	UPAs	Área	Plantas
1	300	+	349	275.097,1	195.866.224	349	275.097,1	195.866.224	100,0	100,0	100,0
2	295	300	15	4.483,2	1.448.355	2	596,1	225.000	13,3	13,3	15,5
3	290	295	4	1.170,3	325.629	2	583,2	105.000	50,0	49,8	32,2
4	285	290	9	2.592,8	748.783	2	574,8	162.315	22,2	22,2	21,7
5	280	285	7	1.977,1	575.162	2	565,0	182.700	28,6	28,6	31,8
6	275	280	10	2.785,2	796.139	2	555,7	140.080	20,0	20,0	17,6
7	270	275	11	2.998,1	959.884	2	543,9	170.639	18,2	18,1	17,8
8	265	270	15	4.021,6	1.267.098	2	534,2	129.000	13,3	13,3	10,2
9	260	265	10	2.627,1	794.799	2	527,5	223.510	20,0	20,1	28,1
10	255	260	14	3.622,5	1.746.130	2	515,2	230.000	14,3	14,2	13,2
11	250	255	10	2.535,6	1.023.454	2	509,4	175.165	20,0	20,1	17,1
12	245	250	7	1.732,9	560.661	2	492,7	160.000	28,6	28,4	28,5
13	240	245	12	2.908,4	1.051.620	2	486,4	163.000	16,7	16,7	15,5
14	235	240	20	4.769,4	1.531.134	2	477,0	162.000	10,0	10,0	10,6
15	230	235	10	2.329,3	957.631	2	466,0	146.844	20,0	20,0	15,3
16	225	230	15	3.412,7	1.041.706	2	458,5	179.000	13,3	13,4	17,2
17	220	225	17	3.783,0	1.287.744	2	445,5	212.840	11,8	11,8	16,5
18	215	220	20	4.359,3	1.514.470	2	435,8	150.000	10,0	10,0	9,9
19	210	215	11	2.333,3	664.670	2	424,3	117.000	18,2	18,2	17,6
20	205	210	15	3.120,1	1.096.432	2	415,1	132.000	13,3	13,3	12,0
21	200	205	15	3.048,9	989.358	2	406,6	137.000	13,3	13,3	13,8
22	195	200	29	5.759,0	1.765.250	2	400,0	128.000	6,9	6,9	7,3
23	190	195	21	4.052,1	1.320.433	2	387,6	101.500	9,5	9,6	7,7
24	185	190	22	4.133,6	1.170.361	2	376,1	120.000	9,1	9,1	10,3
25	180	185	20	3.663,3	1.367.264	2	368,9	125.000	10,0	10,1	9,1
26	175	180	20	3.569,2	975.691	2	355,3	94.000	10,0	10,0	9,6
27	170	175	24	4.142,6	1.350.420	2	345,5	103.382	8,3	8,3	7,7
28	165	170	34	5.731,3	1.863.826	2	339,0	106.000	5,9	5,9	5,7
29	160	165	26	4.231,3	1.373.378	2	327,6	111.000	7,7	7,7	8,1
30	155	160	31	4.902,9	1.636.413	2	313,2	92.495	6,5	6,4	5,7
31	150	155	31	4.745,1	1.620.502	2	305,4	89.000	6,5	6,4	5,5
32	145	150	48	7.099,3	2.177.125	2	294,5	68.000	4,2	4,1	3,1
33	140	145	21	2.997,6	912.631	2	285,0	89.200	9,5	9,5	9,8
34	135	140	44	6.082,6	1.848.641	2	277,0	50.000	4,5	4,6	2,7
35	130	135	42	5.577,6	1.757.811	2	268,0	48.000	4,8	4,8	2,7
36	125	130	40	5.120,3	1.606.745	2	255,5	87.000	5,0	5,0	5,4
37	120	125	51	6.225,1	1.769.865	2	243,5	92.000	3,9	3,9	5,2
38	115	120	55	6.486,2	2.393.930	2	235,0	82.860	3,6	3,6	3,5
39	110	115	34	3.823,4	1.483.451	2	228,0	68.000	5,9	6,0	4,6
40	105	110	54	5.833,3	1.801.057	2	220,0	88.000	3,7	3,8	4,9

Fonte: Elaborada pelos autores com base em Torres et al. (2009).

TABELA 2 - Esquema Amostral para Previsão e Estimativa de Safra de Laranja, Estado de São Paulo, Safra Agrícola 2007/08

(conclusão)

Estrato	Limites		População			Amostra			% Amostra / População		
	Inferior	Superior	Número de UPAs	Área de laranja (ha)	Número de plantas	Número de UPAs	Área de laranja (ha)	Número de plantas	UPAs	Área	Plantas
41	100	105	56	5.751,9	1.865.859	2	206,6	76.654	3,6	3,6	4,1
42	95	100	92	9.004,8	3.763.369	2	193,3	56.175	2,2	2,1	1,5
43	90	95	64	5.926,7	2.091.586	2	184,3	71.785	3,1	3,1	3,4
44	85	90	75	6.600,8	2.116.445	2	177,0	43.000	2,7	2,7	2,0
45	80	85	99	8.224,9	3.130.498	2	166,8	41.000	2,0	2,0	1,3
46	75	80	107	8.335,9	2.752.167	2	159,9	63.133	1,9	1,9	2,3
47	70	75	107	7.775,1	2.636.222	2	146,0	62.000	1,9	1,9	2,4
48	65	70	133	9.057,3	3.238.045	2	137,8	57.530	1,5	1,5	1,8
49	60	65	136	8.506,1	2.613.957	2	126,8	41.637	1,5	1,5	1,6
50	55	60	184	10.658,4	3.385.967	2	119,9	30.000	1,1	1,1	0,9
51	50	55	194	10.242,7	3.701.849	2	105,2	34.500	1,0	1,0	0,9
52	45	50	302	14.436,2	4.650.489	2	95,6	31.500	0,7	0,7	0,7
53	40	45	313	13.347,9	4.058.581	2	87,3	25.750	0,6	0,7	0,6
54	35	40	395	14.861,9	4.558.818	2	75,3	21.352	0,5	0,5	0,5
55	30	35	530	17.328,7	6.040.224	2	61,9	17.900	0,4	0,4	0,3
56	25	30	681	18.928,4	6.155.373	2	57,0	21.000	0,3	0,3	0,3
57	20	25	1.151	26.153,8	8.448.498	2	45,2	14.200	0,2	0,2	0,2
58	15	20	1.656	29.269,8	9.365.972	2	35,4	10.200	0,1	0,1	0,1
59	10	15	2.455	30.794,0	10.508.032	2	25,6	8.700	0,1	0,1	0,1
60	9	10	857	8.309,8	2.919.272	2	18,8	5.670	0,2	0,2	0,2
61	8	9	684	5.895,7	2.019.799	2	17,5	6.570	0,3	0,3	0,3
62	7	8	837	6.310,6	2.419.614	2	15,8	5.200	0,2	0,3	0,2
63	6	7	696	4.666,9	1.551.607	2	12,7	2.850	0,3	0,3	0,2
64	5	6	839	4.795,8	1.900.608	2	11,7	4.100	0,2	0,2	0,2
65	4	5	1.186	5.579,4	1.994.154	2	9,8	3.200	0,2	0,2	0,2
66	3	4	991	3.621,7	1.327.348	2	7,5	3.900	0,2	0,2	0,3
67	2	3	1.228	3.195,5	1.198.620	2	6,0	2.550	0,2	0,2	0,2
68	1	2	1.029	1.691,8	769.778	2	3,5	1.200	0,2	0,2	0,2
69	0	1	2.000	1.013,7	463.836	2	0,2	60	0,1	0,0	0,0
Estado			20.320	730.170	348.088.464	485	293.246	201.676.070	2,4	40,2	57,9

Fonte: Elaborada pelos autores com base em Torres et al. (2009).

TABELA 3 - Número de Upas com Falta de Resposta no Levantamento Amostral para Previsão e Estimativa de Safra de Laranja, Estado de São Paulo, Safra Agrícola 2007/08 (Safra Industrial 2008/09)

Estrato	Produção	Número de plantas em produção	Área em produção
1	107	106	105
2	1	1	1
4	1	1	1
10	1	1	1
15	2	2	2
18	1	1	1
19	1	1	1
22	1	1	1
25	1	1	1
36	2	2	2
37	1	1	1
39	1	1	1
69	1	1	1
Soma	121	120	119
Percentual no estrato 1	88,4	88,3	88,2

Fonte: Dados da pesquisa.

empenho dos técnicos envolvidos e dos diversos retornos às UPAs, os resultados não puderam ser validados devido à recusa de 28,7% delas. Foram empregadas técnicas de imputação de dados. Estas, porém, não se mostraram viáveis, dada principalmente a diversidade dessas unidades produtoras em relação às respondentes e a importância na produção de laranja para a indústria paulista, em função da maioria pertencer às indústrias e aos produtores com contrato de exclusividade.

Por esse motivo, as estatísticas divulgadas na época pelo IEA continuaram a provir do levantamento subjetivo (CASER et al., 2009).

Um dos motivos pelos quais esse tipo de levantamento subjetivo encontra tantos adeptos, principalmente entre os não cientistas, é a grande facilidade para manipulação de tais informações[25]. Enquanto nos levantamentos de UPAs, censitários ou por amostragem, a pressão para forjar dados numa dada UPA não altera muito a estimativa final e, além disso, pode ser facilmente descoberta com técnicas estatísticas de controle de qualidade, nos levantamentos em que a unidade de observação é um município inteiro, tal pressão para modificar dados pode ter resultados proporcionalmente grandes. Dessa forma, embora os levantamentos subjetivos não impliquem necessariamente em fraudes, elas são enormemente facilitadas. Com estatísticas agrícolas ruins, ganham somente os especuladores de toda sorte; portanto, mais ainda se fazem necessários o rigor científico e a seriedade de tais serviços.

Numa tentativa de salvar o levantamento e tirar dele um mínimo de resultados, não apenas em respeito aos produtores alocados na amostra que responderam, mas também para evitar desperdício de dinheiro público, utilizaram-se algumas técnicas de tratamento de falta de resposta. O teste de Little foi aplicado aos dados de produção, número de plantas e área plantada, provenientes do levantamento de campo por amostragem, mais a área total de laranja proveniente do censo agropecuário paulista (projeto LUPA), sendo as três primeiras variáveis com valores perdidos e a última sem valores perdidos. O resultado mostra um qui-quadrado igual a 63,588, com 9 graus de liberdade e p-valor < 0,001. Portanto, no nível de significância de 5%, rejeita-se a hipótese de que a falta de resposta seja completamente aleatória (*missing completely at random - MCAR*) ou que os dados perdidos sejam completamente ao acaso. Em outras palavras, a falta de resposta de produção, número de plantas e área plantada, provenientes do levantamento de campo por amostragem, não é inde-

[25]O que não significa, necessariamente, que este ou aquele levantamento subjetivo tenha sido manipulado. Chama-se a atenção neste artigo apenas para a possibilidade potencial de manipulação nesse tipo de levantamento, ademais utilizado também por outras instituições, sem denunciar especificamente que qualquer deles tenha sido manipulado.

pendente da área total de laranja proveniente do censo agropecuário paulista (projeto LUPA), que foi utilizada na estratificação da amostra. De fato, a falta de resposta depende do tamanho da exploração citrícola, o que faz sentido, uma vez que as maiores são as de maior produtividade e recursos tecnológicos.

Usando uma das técnicas ingênuas (*naïve*), a saber, a eliminação de casos completos (*listwise delection*), foi possível obter estimativas. Os resultados obtidos somente para os citricultores que colaboraram (como a produção, da ordem de 244 milhões de caixas) são viesados, provavelmente para baixo dos valores verdadeiros, uma vez que as UPAs com área de laranja acima de 300 ha, que não responderam, são responsáveis por grandes produções, em grandes áreas, com adensamento de plantio e alta produtividade (Tabela 4). A imputação condicional pela média forneceu valor intermediário (328 milhões de caixas). Também foram feitos cálculos utilizando técnicas múltiplas de imputação de dados, obtendo-se estimativas aparentemente altas (457 a 478 milhões de caixas). Para efeito de comparação, o levantamento subjetivo resultou no valor de 366 milhões de caixas, enquanto a opinião de um especialista indicava 334 milhões de caixas[26]. A não convergência dos resultados inviabiliza, da mesma forma como os outros métodos, sua aplicação e a tentativa em se obter uma estimativa aderente à realidade.

Levantamentos estatísticos do tipo aqui analisado baseiam-se em relação de mútua confiança entre os que solicitam e os que cedem os dados: uma vez que a produção não é medida no campo, nem as plantas em produção são contadas, os responsáveis pelo levantamento devem confiar que os dados fornecidos pelo produtor estejam corretos; por outro lado, o produtor acredita que seus dados individuais não serão divulgados nem usados contra ele sob qualquer pretexto. Em outras palavras, esse tipo de levantamento funciona somente quando há colaboração estreita entre as partes, o que geralmente acontece, mas que se mostrou um problema no presente caso, com a recusa de dados por boa parte dos grandes produtores. Num extremo, pode-se argumentar que o produtor (ou a indústria que controla grande número de produtores) tem direito e liberdade de não responder, pelos motivos que bem entender. O IEA não dispõe de dispositivo legal que obrigue os informantes a fornecer dados, baseando seu trabalho exclusivamente na relação de mútua confiança. Em outro extremo, pode-se argumentar que recursos públicos não devem ser desperdiçados com um setor que se recusa a colaborar. Entretanto, instituições públicas têm responsabilidade e papel relevante na democratização de informações estatísticas (PINO, 1999, 2001b), o que no caso da laranja significa estimativas de produção às quais todos os envolvidos na cadeia de produção tenham acesso. Como a negativa de dados é feita por partes que dispõem de suas próprias estatísticas de produção, eventualmente elas poderiam ser denunciadas por abuso de poder econômico. Além disso, muitas das reclamações a respeito das estatísticas citrícolas partem exatamente dos que costumam se negar a colaborar para sua melhoria. Entretanto, confrontos devem ser evitados e, portanto, uma solução de compromisso[27] deve ser procurada, de tal forma a agradar todas as partes envolvidas.

5 - CONSIDERAÇÕES FINAIS

O esquema amostral testado mostrou ter potencial para obter dados de boa qualidade. O fracasso no campo deveu-se a causas não estatísticas, o que não o inutiliza para futuras aplicações. É do interesse dos setores, tanto público quanto privado, o conhecimento da realidade da produção nacional, urgindo, assim, a colaboração entre esses setores.

Depois da safra aqui retratada, a postura daqueles que recusavam informações vem se alterando positivamente, isto é, cada vez mais colaborando no fornecimento de informações para compor a estimativa de safra agrícola para a laranja. Todavia, ainda há muitos produtores, expressivos comercialmente, relutantes a

[26]De acordo com o Dr. Antonio Ambrosio Amaro, é possível obter essa estimativa analisando a série histórica de produção de laranja, número de plantas por faixa etária, ajustada com informações climáticas, bem como informações obtidas junto ao setor no período inicial de colheita a respeito do diâmetro das frutas, medido por calibrador de frutas (*fruit sizer*) na linha de produção industrial, que permite calcular o número de frutas por caixa (comunicação verbal).

[27]Formalizada ou não na forma de termo de ajustamento de conduta (TAC).

TABELA 4 - Comparação das Estimativas Obtidas por Levantamento Amostral com Estimativas Obtidas por Outros Métodos, Previsão e Estimativa de Safra de Laranja, Estado de São Paulo, Safra Agrícola 2007/08 (Safra Industrial 2008/09)

Levantamento	Método ou região	Produção (caixa 40,8 kg)	Número de plantas em produção	Área em produção (ha)	Produção por planta (cx./pé)	Densidade (planta/ha)
Amostral[1]	Eliminação de casos não disponíveis (*pairwise delection - PD*)	243.788.903	149.393.448	438.664	1,63	341
	Imputação condicional pela média ou imputação por regressão	327.949.197	211.155.090	568.266	1,55	372
	Imputações múltiplas pelo método de Monte Carlo via cadeias de Markov (MCMC)	478.362.468	271.293.683	762.361	1,76	356
	Imputações múltiplas pelo método de regressão	457.389.357	269.750.572	737.881	1,70	366
	Imputações múltiplas com maximização da esperança (*expectation maximization - EM*)	470.763.789	275.615.311	757.860	1,71	364
Censitário (LUPA)[2]						477
Subjetivo[3]	Total do Estado	365,815,444	183,457,158	573,303,6	1,99	320
	Somente zona citrícola	349,918,705	176,273,376	550,854,3	1,99	320
Opinião de especialista[4]		334,000,000				

Fonte: [1]Elaborada com base na pesquisa; [2]Elaborado sobre dados originais do censo agropecuário paulista (projeto LUPA 2007/2008). Os valores apresentados eventualmente diferem daqueles constantes em Torres et al. (2009), uma vez que alguns casos foram eliminados, como pomares domésticos e outros sem interesse para a presente pesquisa; [3]Elaborado com base no Levantamento Subjetivo para Previsão e Estimativa de Safras, Instituto de Economia Agrícola (IEA) e Coordenadoria de Assistência Técnica Integral (CATI); [4]Comunicação verbal do Dr. Antonio Ambrosio Amaro.

participarem, e espera-se que aos poucos eles se conscientizem da responsabilidade em se obter um número oficial mais acurado.

As conclusões aqui apresentadas serviram de base para negociações que resultaram em acordo entre o IEA, a CATI e a CONAB para o levantamento oficial conjunto da safra agrícola 2010/2011, safra industrial 2011/2012 (CAMARGO; FRANCISCO, 2011), bem como em acordo verbal para a colaboração das indústrias.

LITERATURA CITADA

ASSOCIAÇÃO BRASILEIRA DE CITRICULTORES - ASSOCITRUS. Diferenças na estimativa de safra influenciam remuneração do produtor. **Informativo Associtrus**, v. 7, n. 36, p. 3, jun./jul. 2011.

BRANDEN, K. V.; VERBOVEN, S. Robust data imputation. **Computational Biology and Chemistry**, Amsterdam, Vol. 33, Issue 1, pp. 7-13, 2008.

BREIMAN, L. et al. **Classification and regression trees**. Belmont: Wadsworth International Group. 1984.

BRUNI, R. Discrete models for data imputation. **Discrete Applied Mathematics**, Amsterdam, Vol. 144, Issues 1-2, pp. 59-69, 2004.

BUCK, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. **Journal of the Royal Statistical Society:** Series B, Hoboken, Vol. 22, Issue 2, pp. 302-306, 1960.

CAMARGO, F. P.; FRANCISCO, V. L.F.S. Estimativa de safra de laranja no estado de São Paulo. **Informações Econômicas**, São Paulo, v. 41, n. 5, 2011.

CAMPOS, H.; PIVA, L. H. O. Dimensionamento de amostra para estimativa e previsão de safra no estado de São Paulo. **Agricultura em São Paulo**, São Paulo, v. 31, t. 3, p. 65-88, 1974.

CASER, D. V. et al. Previsão da safra agrícola 2007/08 para a cultura da laranja. **Análises e Indicadores do Agronegócio**, São Paulo, v. 4, n. 1, jan. 2009. Disponível em <http://www.iea.sp.gov.br/out/verTexto.php?codTexto=9792>. Acesso em: 15 jan. 2009.

DYBOWSKI, R. Classification of incomplete feature vectors by radial basis function networks. **Pattern Recongnition Letters**, Amsterdam, Vol. 19, Issue 14, pp. 1257-1264, 1998.

ENGELS, J. M.; DIEHR, P. Imputation of missing longitudinal data: a comparison of methods. **Journal of Clinical Epidemiology**, Amsterdam, Vol. 56, Issue 10, pp. 968-976, 2003.

FAGUNDES, P. R. S. et al. Cultura da laranja no estado de São Paulo. **Informações Econômicas**, v. 40, n. 9, p. 54-67, 2010.

FRANCISCO, V. L. F. S. et al. Controle de qualidade de dados estatísticos: o levantamento censitário de unidades de produção agrícola. **Agricultura em São Paulo**, São Paulo, v. 45, n.1, p.33-58, 1998.

_____. et al. Modelo estatístico e econômico para estimativa da safra brasileira de café. **Informações Econômicas**, São Paulo, v. 40, n. 12, p. 26-36, 2010.

HAIR JUNIOR, J. F. et al. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005. 593 p.

IACUS, S. M.; PORRO, G. Missing data imputation, matching and other applications of random recursive partitioning. **Computational Statistics & Data Analysis**, Amsterdam, Vol. 52, Issue 2, pp. 773-789, 2007.

INSTITUTO DE ECONOMIA AGRÍCOLA - IEA. Mário Zaroni (1921-2003). **Agricultura em São Paulo**, SP, São Paulo, v.25, t. 1-2, p.323-325, 1978.

_____. Salomão Schattan (1907-1975). **Agricultura em São Paulo**, SP, São Paulo, v. 50, t. 2, p. 111-116, 2003.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, Amsterdam, Vol. 38, Issue 18, pp. 2895-2907, 2004.

KIM, K. Y.; KIM, B. J; YI, G. S. Reuse of imputed data in microarray analysis increases imputation efficiency. **BMC Bioinformatics**, London, Vol. 5, Issue 160, pp. 1-9, 2004.

KISH, L. **Survey sampling**. New York: Wiley, 1965. 643 p.

LITTLE, R. J. A. A test of missing completely at random for multivariate data with missing values. **Journal of the American Statistical Association**, v83, p. 1198-1202, 1988.

LIU, C. Missing data imputation using the multivariate **t** distribution. **Journal of Multivariate Analysis**, Amsterdam, Vol. 53, Issue 1, pp. 139-158, 1995.

MARTINS, V. A. **Amostragem probabilística e imagens de satélite para estimativa de área de citros**. 2009. 155

p. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2009.

MOJIRSHEIBANI, M. The Glivenko-Cantelli theorem based on data with randomly imputed missing values. **Statistics & Probability**, Amsterdam, Vol. 55, Issue 4, pp. 385-396, 2001.

MORICOCHI, L.; PINO, F. A.; VEGRO, C. L. R. Método Delphi como alternativa para previsão de safras: o exemplo do café. **Informações Econômicas**, São Paulo, v. 25, n. 12, p. 47-52, dez. 1995.

NIELSEN, S. F. Nonparametric conditional mean imputation. **Journal of Statistical Planning and Inference**, Amsterdam, Vol. 99, Issue 2, pp. 129-150, Dez. 2001.

OBA, S. et al. A Bayesian missing value estimation method for gene expression profile data. **Bioinformatics**, Oxford, Vol. 19, Issue 16, pp. 2088-2096, 2003.

OLINSKY, A.; CHEN, S.; HARLOW, L. The comparative efficacy of imputation methods for missing data in structural equation modeling. **European Journal of Operational Research**, Amsterdam, Vol. 151, Issue 1, pp. 53-79, 2003.

PINO, F. A. Análise do viés em alguns procedimentos para falta de resposta e para erros de resposta em levantamentos por amostragem. **Agricultura em São Paulo**, São Paulo, v. 36, n. 2, p. 147-153, 1989.

_____. Centenário do censo agronômico. **Informações Econômicas**, São Paulo, v. 35, n. 5, p. 85-97, maio 2005.

_____. Detecção e correção de erros em levantamentos agrícolas. **Pesquisa Agropecuária Brasileira**, Brasília, v. 21, n. 9, p. 979-985, set. 1986.

_____. Estatísticas agrícolas para o século XXI. **Agricultura em São Paulo**, São Paulo, v. 46, n. 2, p. 71-105, 1999.

_____. Estimação subjetiva de safras agrícolas. **Informações Econômicas**, São Paulo, v. 31, n. 6, p. 55-58, jun. 2001a.

_____. Meio século de amostragem nas estatísticas agrícolas. São Paulo: IEA, 2004. (Seção Políticas Públicas). Disponível em: <http://www.iea.sp.gov.br/out/LerTexto.php?codTexto=1236>. Acesso em: 18 fev. 2004.

_____. Previsão de custo de levantamento estatístico por amostragem. **Informações Econômicas**, São Paulo, v. 32, n. 7, p. 55-59, jul. 2002.

_____. Projeto LUPA: uma odisséia. **Informações Econômicas**, São Paulo, v. 30, n. 11, p. 65-68, nov. 2000.

_____. Quem tem medo do censo agropecuário? **Análise e Indicadores do Agronegócio**, São Paulo, v. 1, n. 3, mar. 2006. Disponível em: < http://www.iea.sp.gov.br/out/LerTexto.php?codTexto=4860>. Acesso em: 09 mar. 2006.

_____. Tendências em informações agropecuárias. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 10. Foz do Iguaçu, 2001. **Anais...** São José dos Campos: INPE, 2001. Imagem Multimídia. 2001b. CD-ROM.

_____. AMARO, A. A. Previsão de safras de citros: algumas possibilidades no estado de São Paulo. **Laranja**, Cordeirópolis, v. 7, n. 2, p. 403-422, 1986.

_____. CASER, D. V. **Falta de respostas em levantamentos por amostragem**: um estudo de caso. São Paulo: IEA, 1984. 25 p. (Relatório de Pesquisa, 8).

_____. CÉZAR, S. A. S. G.; AMARO, A. A. Modelling and forecasting Florida orange production. In: INTERNATIONAL CITRUS CONGRESS, 7., Acireale, Mar. 1992. **Proceedings...** Acireale: International Society of Citriculture,

1992. p. 1199-1200.

PINO, F. A.; FRANCISCO, V. L. F. S. Controle de qualidade em levantamento agrícola por amostragem em São Paulo. **Informações Econômicas**, São Paulo, v. 31, n. 6, p. 7-24, jun. 2001.

_____; _____; LORENA NETO, B. Previsão e estimativa de safras cafeeiras no Estado de São Paulo. **Agricultura em São Paulo**, São Paulo, v. 48, t. 1, p. 57-68, 2001.

_____. et al. (Orgs.) **Levantamento censitário de unidades de produção agrícola do Estado de São Paulo**. São Paulo: IEA/CATI/SAA, 1997. 4 v.

PIVA, L. H. O. et al. Avicultura na economia agrícola de São Paulo. **Agricultura em São Paulo**, São Paulo, v. 22, t. 1-2, p.305-340. 1975.

QIN, Y. et al. POP algorithm: kernel-based imputation to treat missing values in knowledge discovery from databases. **Expert Systems with Applications**, Amsterdam, Vol. 36, Issue 2, pp. 2794-2804, 2009.

SANTOS, V. L. F. et al. Dimensionamento de amostra para levantamento da citricultura paulista. **Pesquisa Agropecuária Brasileira**, Brasília, v. 11, n. 1, p. 15-21, 1987.

SCHATTAN, S. Obtenção de estatísticas agrícolas pelo método de amostragem: experiências visando a criação de uma organização permanente. **Agricultura em São Paulo**, São Paulo, v. 50, t. 2, p.81-109, 2003.

SENTAS, P.; ANGELIS, L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. **The Journal of Systems and Software**, Vol. 79, Issue 3, pp. 404-414, 2006.

SONG, Q.; SHEPPERD, M. A new imputation method for small software project data sets. **The Journal of Systems and Software**, Amsterdam, Vol. 80, Issue 1, pp. 51-62, 2007.

TORRES, A. J. et al. (Org.). **Projeto LUPA 2007/08:** censo agropecuário do Estado de São Paulo. São Paulo: IEA/CATI/SAA, 2009. 381 p.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. **Bioinformatics**, Oxford, Vol. 17, Issue 6, pp. 520-525, 2001.

TWISK, J.; VENTE, W. Attrition in longitudinal studies: how to deal with missing data. **Journal of Clinical Epidemiology**, Amsterdam, Vol. 55, Issue 4, pp. 329-337, 2002.

VERBOVEN, S.; BRANDEN, K. V.; GOOS, P. Sequential imputation for missing values. **Computational Biology and Chemistry**, Amsterdam, Vol. 31, Issues 5-6, pp. 320-327, 2007.

WASITO, I.; MIRKIN, B. Nearest neighbours in least-squares data imputation algorithms with different missing patterns. **Computational Statistics & Data Analysis**, Amsterdam, Vol. 50, Issue 4, pp. 926-949, 2006.

ESTIMATIVA DE SAFRA DE LARANJA EM 2008: um suco amargo

RESUMO: *Uma amostra probabilística, estratificada pelo tamanho da cultura em cada unidade de produção, foi calculada e aplicada no campo em 2008 a fim de estimar a produção de laranja no Estado de São Paulo. Entretanto, 25% dos produtores amostrados recusaram-se a responder, destruindo o esquema amostral e arruinando as estimativas. A falta de resposta foi atribuída a segredo empresarial e*

cláusulas contratuais, uma vez que os não respondentes concentravam-se em unidades produtoras pertencentes a (ou contratadas por) indústrias citrícolas. Apresenta-se um levantamento sobre o estado da arte a respeito da questão difundida de valores perdidos e imputação de dados e discutem-se estratégias para o futuro da estimação da safra de laranja. Depois da safra aqui discutida, o comportamento dos não respondentes tem mudado progressivamente.

Palavras-chave: *teste de metodologia, controle de qualidade de dados de levantamento, valores perdidos, imputação de dados, revisão de dados de levantamento.*

ORANGE CROP ESTIMATION IN 2008:
a bitter juice

ABSTRACT: *A probabilistic sample, stratified by crop size on each farm, was calculated and applied in the field in order to estimate orange production in Sao Paulo state, Brazil. 25% of the sampled farmers refused to answer, destroying the sample design and ruining the estimation. Non-response was attributed to trade secrecy and contractual restrictions, since non-respondents were concentrated on farms owned (or contracted by) citrus agri-business. A state-of-the-art survey on the pervasive question of missing values and data imputation is presented, and the strategies for future orange crop estimates are discussed. Since the 2008 crop discussed here, non-respondents' behavior has shown progressive signs of change.*

Key-words: *methodology test, quality control of survey data, missing values, data imputation, editing of survey data.*

Recebido em 30/06/2011. Liberado para publicação em 03/08/2011.