Working Paper nº 48

Estimação em pequenos domínios

Pedro Simões Coelho

Working Paper nº 48

ISSN: 0872-895X

Depósito Legal nº: 90631/95

Pedro Simões Miguel

Abril 1996

Estimação em pequenos domínios

Pedro Coelho

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Abstract

This paper approaches some issues related to domain estimation in the frame of a survey. The survey design stage is approached, emphasising that the problem should be viewed in an overall perspective. Stress is, nevertheless, placed at the estimation stage, through the highlight of estimation methods, with references to their uses and relative merits. A special attention is also paid to the problem of small domain estimation, where sample size is seldom sufficient to obtain acceptable direct estimates.

Resumo

Neste artigo pretende-se apresentar algumas questões relacionadas com a estimação em domínios no âmbito de sondagens. São abordadas as etapas de planeamento e desenho da sondagem, sublinhando que o problema deve ser visto numa perspectiva global. O ênfase é no entanto colocado na etapa de estimação, através de uma identificação abrangente de métodos de estimação, com referências às suas aplicações e qualidades relativas. É dada uma atenção especial ao problema de estimação em pequenos domínios, onde a dimensão das amostras é normalmente insuficiente para obter estimativas directas de precisão aceitável.

Palavras chave: sondagens, domínios, pequenos domínios, estimação, *design based*, *model based*, *model assisted*.

1. Introdução

As sondagens, inicialmente pensadas para fornecer informação ao nível de grandes áreas, são cada vez mais utilizadas para fornecer estimativas para domínios de diversa natureza e dimensão.

A necessidade de produção de estimativas fiáveis para o total de variáveis de interesse em pequenos domínios (e em particular pequenas áreas) no âmbito das sondagens tem vindo a tornar-se mais premente nos últimos anos. Devido a esta procura crescente de estatísticas fiáveis ao nível de pequenos domínios, quer pelo sector público quer pelo sector privado, a estimação em pequenos domínios tem vindo a ganhar uma importância crescente no âmbito da teoria das sondagens.

Na produção de tais estimativas, emerge um problema evidente, sobretudo em domínios de muito pequena dimensão, já que as dimensões amostrais no interior de tais domínios são habitualmente demasiado pequenas para que se possam utilizar estimadores directos[1].

Como resultado destas preocupações recentes, têm vindo a ser propostos vários tipos de estimadores que combinem informação auxiliar com os dados da sondagem para os domínios de estudo.

Embora este artigo se oriente fundamentalmente para os problemas de estimação, dever-se-á sublinhar que os problemas associados aos pequenos domínios devem ser vistos numa perspectiva global e considerados na fase de planeamento da sondagem. O desenho da sondagem deverá assim, sempre que possível, contemplar este fenómeno de forma a permitir a produção de dados fiáveis para os domínios, remetendo a manipulação do problema ao nível da estimação para uma solução de último recurso.

2. Abordagem sistémica

Muitas vezes, os problemas que surgem na estimação para domínios resultam da não consideração dessas necessidades no início do processo de planeamento da sondagem.

De facto, já foi referido que estes não devem ser exclusivamente encarados como problemas de estimação, sendo antes desejável a definição de uma estratégia global que envolva a preocupação de dar resposta a estas necessidades nas fases de planeamento, desenho e estimação.

Tal passa obviamente por uma identificação prévia de todos os domínios para os quais se pretende obter informação, sendo apenas para os domínios não planeados ou para

[1] Trata-se de estimadores desenvolvidos numa óptica *design based*. Uma classificação de estimadores, onde este conceito é precisado, é apresentada no capítulo 4.

os domínios planeados referentes a populações raras que se deverão desenvolver métodos de estimação especiais.

Esta consideração motiva uma primeira classificação dos domínios.

Domínios planeados: domínios que são identificados na fase de planeamento da sondagem; tipicamente, tratam-se de estratos ou grupos de estratos para os quais podem ser planeadas as amostras desejadas.

Domínios não planeados: domínios que não foram identificados na fase de planeamento da sondagem e que podem atraversar os estratos definidos no desenho.

3. Considerações ao nível do desenho

Habitualmente são introduzidos vários níveis de compromisso nas várias etapas da amostragem e da recolha de dados a fim de satisfazer restrições teóricas e operacionais. As estimativas para domínios deverão igualmente fazer parte deste compromisso.

Há duas formas fundamentais de tomar em consideração os pequenos domínios ao nível do desenho: através da estratégia de alocação e do número de etapas (*clustering*).

3.1. *Estratégia de alocação*

Em geral, uma estratégia de alocação óptima ao nível global reparte as amostras das unidades primárias de forma aproximadamente proporcional à sua dimensão. A fiabilidade das estimativas para as unidades de menor dimensão é como tal prejudicada.

Do ponto de vista da estimação em domínios, será preferível uma alocação de compromisso, i.e., produzindo pequenas reduções nas dimensões das amostras para as unidades de maior dimensão que são transferidas para as unidades de menor dimensão. Tal tem normalmente pouco efeito na fiabilidade dos dados ao nível das grandes unidades e ao nível global, mas poderá corresponder a ganhos muito significativos nas unidades de menor dimensão.

3.2. *Número de etapas*

As sondagens de larga escala envolvem normalmente desenhos multietápicos, com unidades primárias relativamente grandes, de forma a reduzir os custos envolvidos.

No entanto, a utilização de desenhos com muitas etapas não se mostra adequada para a estimação em pequenos domínios, já que nesse caso será provável que alguns tenham amostras apreciáveis, enquanto que outros poderão apresentar amostras de dimensão muito reduzida ou mesmo nula.

Como tal, deverão ser realizados esforços para reduzir o grau de *clustering* na amostragem. Este esforço será normalmente associado à substituição de estratos de grande dimensão por vários de menor dimensão, com o objectivo de fazer com que os domínios contenham essencialmente estratos completos, tornando desta forma mais estável a dimensão da amostra.

4. Questões ligadas à estimação

Não obstante a atenção que deverá ser dada à informação dos domínios nas primeiras fases da sondagem, é a estimação que mais tem interessado os investigadores.

De facto, mesmo levando a cabo um planeamento rigoroso das necessidades ao nível dos domínios, existirão sempre múltiplas situações onde, por via da impossibilidade da identificação prévia de domínios de interesse ou por via da sua pequena dimensão associada a questões de restrição orçamental, se torna necessário recorrer a métodos de estimação especiais.

É largamente reconhecido que as estimativas directas para pequenos domínios poderão apresentar variâncias de dimensão inaceitável devido às pequenas dimensões amostrais nos domínios. Tal conduz à necessidade de recorrer à informação de outros domínios relacionados, com o objectivo de obter estimativas mais precisas.

Será então útil começar por apresentar uma classificação dos vários tipos de estimadores disponíveis. Estes estimadores seguem uma classificação que está intimamente ligada a duas escolas de pensamento dentro da estatística que serão daqui por diante designadas por *design based* e *model based*[2].

- *Design based* { Estimadores directos; Estimadores directos modificados }
- *Model based* { Estimadores sintéticos }
- *Model assisted* { Estimadores combinados }

[2] Uma referência mais detalhada sobre este assunto pode ser encontrada em Särndal, Swensson e Wretman (1992).

Esta classificação segue de perto a proposta por Singh, Gambino e Mantel (1994), pretendendo-se no entanto aqui isolar (ao contrário dos referidos autores) os estimadores combinados relativamente à classe de estimadores *model based*.

4.1. *Estimadores directos*

São estimadores que utilizam valores da variável de estudo só para o período de tempo de interesse e apenas para as unidades do domínio. Estes estimadores podem usar informação de variáveis auxiliares dentro ou fora do domínio.

A este nível procura-se produzir estimadores que apresentem propriedades estatísticas interessantes do ponto de vista do desenho, e em particular o não enviesamento.

O estimador mais simples é o estimador π ou estimador de Horvitz-Thompson

$$\hat{\tau}_{d\pi} = \sum_{i \in s_d} \pi_i^{-1} y_i \ , \qquad \textbf{eq. 4.1}$$

onde

s_d : parte da amostra no domínio d;
π_i : probabilidade de inclusão da unidade i na amostra

Como é bem sabido, este estimador é centrado, mas poderá ter elevada variância devido à aleatoriedade da dimensão da amostra no domínio. Repare-se, no entanto, na facilidade da sua aplicação, já que este mantém a forma habitual tal como se uma amostra de dimensão n_d fosse seleccionada directamente a partir da sub-população que constitui o domínio.

No caso em que a dimensão do domínio N_d é conhecida, recorre-se ao estimador pós-estratificado

$$\hat{\tau}_{d,pst} = N_d \frac{\sum_{i \in s_d} \pi_i^{-1} y_i}{\sum_{i \in s_d} \pi_i^{-1}} = N_d \frac{\hat{\tau}_{d\pi}}{\hat{N}_{d\pi}} . \qquad \textbf{eq. 4.2}$$

Este estimador utiliza informação auxiliar referente à dimensão do domínio apresentando-se como tal mais estável e habitualmente de maior qualidade do que o estimador anterior[3].

---

[3] Repare-se a título de exemplo que numa sondagem aleatória simples a eficiência relativa dos estimadores virá: $V(\hat{\tau}_{d\pi}) / AV(\hat{\tau}_{d,pst}) = 1 + (1 - P_d) / (CV_{yd})$. Onde CV_{yd} é o coeficiente de variação da variável de interesse no domínio d e P_d a dimensão relativa do domínio. A vantagem do estimador pós-estratificado torna-se evidente principalmente em populações raras.

Tratando-se da estimação de um rácio, resultam todas as respectivas propriedades, verificando-se nomeadamente que o estimador permanece aproximadamente centrado.

No caso de uma sondagem estratificada (pós-estratificada), em que a dimensão do domínio em cada estrato (pós-estrato) seja conhecida, tem-se um estimador pós-estratificado alternativo

$$\hat{\tau}_{d,st,pst} = \sum_h N_{h,d} \frac{\sum_{i \in s_{h,d}} \pi_i^{-1} y_i}{\sum_{i \in s_{h,d}} \pi_i^{-1}} . \qquad \textbf{eq. 4.3}$$

No fundo, o total do domínio em cada um dos estratos (pós-estratos) será estimado separadamente.

Claro que este estimador continua a manifestar alguma instabilidade, já que os n_{hd} podem ser pequenos em valor esperado, permanecendo, no entanto, aproximadamente centrado.

Será no entanto possível obter estimadores de melhor qualidade, recorrendo a informação auxiliar (para além da dimensão dos domínios).

O recurso a informação auxiliar torna-se aliás de particular importância neste contexto, de forma a compensar a escassez dos dados da amostra em pequenos domínios.

Neste sentido, é possível utilizar um estimador por rácio (ou quociente) semelhante em forma ao estimador pós-estratificado, mas utilizando uma variável auxiliar X em vez das dimensões dos domínios.

Tem-se assim os estimadores,

$$\hat{\tau}_{d,r} = \tau_{xd} \hat{R}_d = \tau_{xd} \frac{\hat{\tau}_{d\pi}}{\hat{\tau}_{xd\pi}} \qquad \textbf{eq. 4.4}$$

e

$$\hat{\tau}_{d,st,r} = \sum_h \tau_{xh,d} \hat{R}_{h,d} = \sum_h \tau_{xh,d} \frac{\hat{\tau}_{h,d\pi}}{\hat{\tau}_{xh,d\pi}}, \qquad \textbf{eq. 4.5}$$

de onde se verifica que o estimador pós-estratificado acaba por aparecer como um caso particular do estimador por rácio em que a variavel auxiliar é a variável indicatriz de pertença ao domínio.

A motivação que suporta estes estimadores resulta da expectativa de uma proporcionalidade aproximada entre a variável de interesse e variável auxiliar.

Finalmente, propõe-se o estimador pela regressão. Trata-se de um estimador mais genérico que procura ter em conta as diferenças entre as sub-populações dos domínios e os valores gerados pela amostra de um vector de variáveis auxiliares. A construção do estimador passa pelo ajustamento de um modelo de regressão.

O estimador tem genericamente a forma

$$\hat{\tau}_{d,reg} = \hat{\tau}_d + \left(\tau_{xd} - \hat{\tau}_{xd}\right)\hat{B}_d, \qquad \textbf{eq. 4.6}$$

onde $\hat{\tau}_d$ é um estimador do total da variável de interesse no domínio, habitualmente o estimador π ou o estimador pós-estratificado, e $\hat{\tau}_{xd}$ tem o mesmo significado relativamente ao vector de variáveis auxiliares.

Os $\hat{B}_d$ são os estimadores dos parâmetros da regressão, obtidos a partir dos dados da amostra de cada domínio d por

$$\hat{B}_d = \left[\sum_{i\in s_d} \upsilon_i^{-1}\pi_i^{-1}x_i x_i'\right]^{-1}\sum_{i\in s_d} \upsilon_i^{-1}\pi_i^{-1}x_i y_i\,, \qquad \textbf{eq. 4.7}$$

onde os υ_i são os pesos da regressão, y_i o valor observado da variável de interesse para o individuo i e x_i o vector coluna das J variáveis auxiliares para o individuo i.

Obviamente que o estimador pela regressão pode igualmente ser aplicado no âmbito de uma sondagem estratificada.

O pressuposto que está subjacente a este estimador consiste em considerar que a variável de interesse poderá ser explicada através de um modelo de regressão ξ do tipo

$$\begin{cases} E_\xi(y_i) = x_i'\beta_d \\ V_\xi(y_i) = \sigma_d^2\upsilon_i \end{cases} \qquad \textbf{eq. 4.8}$$

Este estimador apresenta propriedades estatísticas muito interessantes. Na verdade, prova-se que é aproximadamente centrado no desenho (com enviesamento assimptoticamente nulo) independentemente da validade do modelo de regressão linear que está subjacente. Por seu lado, a variância no desenho será tanto menor, quanto mais a verdadeira relação entre a variável de interesse e as variáveis auxiliares se aproximar da situação de relação linear perfeita sugerida pelo modelo.

Trata-se de um estimador bastante analisado na bibliografia. Uma análise detalhada pode ser encontrada em Cassel, Särndal eWretman (1976), Särndal (1980,1982), Isaki e Fuller (1982), Wright (1983), Särndal, Swensson e Wretman (1992).

Se o domínio d tiver uma dimensão razoável, o modelo gera o estimador pela regressão de boa qualidade. Os problemas levantam-se quando existe um grande numero de domínios de pequena dimensão. Neste caso poderão resultar estimativas fracas para os parâmetros β_d, conduzindo à necessidade de recorrer a outro tipo de estimadores, nomeadamente os indirectos analisados na secção seguinte.

Note-se ainda que os estimadores apresentados anteriormente podem ser encarados como casos particulares do estimador pela regressão. Na verdade, os estimadores pós-estratificado e por rácio baseiam-se implicitamente em modelos do tipo,

$$\begin{cases} E_\xi(y_i) = \beta_d \\ V_\xi(y_i) = \sigma_d^2 \end{cases} \quad \text{e} \quad \begin{cases} E_\xi(y_i) = \beta_d x_i \\ V_\xi(y_i) = \sigma_d^2 x_i \end{cases}, \text{ respectivamente;}$$

onde x_i é um escalar.

4.2. *Estimadores directos modificados*

Estes estimadores podem usar informação de outros domínios quer para a variável de interesse, quer para as variáveis auxiliares, mantendo ainda propriedades interessantes do ponto de vista do desenho, nomeadamente o não enviesamento aproximado.

Um estimador modificado pode ser encarado como um estimador directo com um ajustamento sintético para o enviesamento do modelo. Como o ajustamento terá valor esperado aproximadamente nulo relativamente ao desenho, o estimador modificado é aproximadamente centrado no desenho se o estimador directo o for.

Um exemplo para um estimador modificado, baseado no estimador pela regressão vem,

$$\hat{\tau}_{d,sreg} = \hat{\tau}_d + (\tau_{xd} - \hat{\tau}_{xd})\hat{B}. \qquad \textbf{eq. 4.9}$$

Trata-se do estimador pela regressão para o domínio d, onde $\hat{B}_d$ é substituído por um estimador sintético

$$\hat{B} = \left[\sum_{i\in s} \upsilon_i^{-1}\pi_i^{-1}x_i x_i'\right]^{-1} \sum_{i\in s} \upsilon_i^{-1}\pi_i^{-1}x_i y_i . \qquad \textbf{eq. 4.10}$$

Repare-se que $\hat{B}$ é agora estimado a partir da amostra global. O pressuposto implícito é o de que a relação linear entre a variável de interesse e as variáveis auxiliares se faz sentir de forma semelhante no domínio e no conjunto da população.

A grande vantagem reside no facto de $\hat{B}$ ser estimado com base em uma amostra de maior dimensão, reduzindo assim a sua variância e tornando-se mais provável que os dados suportem estimativas estáveis dos parâmetros do modelo. Claro que a escolha final depende da dimensão da variância de $\hat{B}_d$ relativamente à variação dos parâmetros B_d entre os vários domínios.

Este estimador corresponde a uma situação extrema em que os parâmetros do modelo são estimados com base em dados combinados de todos os domínios. Um caso intermédio, que origina igualmente estimadores directos modificados, corresponde à agregação dos domínios num número limitado de grupos de tal forma que os parâmetros do modelo são considerados constantes para todos os elementos de um mesmo grupo de domínios.

Desta forma, é possível reduzir o número de parâmetros a estimar (relativamente ao estimador directo), utilizando no entanto hipóteses menos restritivas dos que as que estão subjacentes ao modelo único para todos os domínios.

Considerando a população segmentada em G grupos de domínios o estimador pela regressão será então suportado por um modelo do tipo

$$\begin{cases} E_\xi(y_i) = x_i'\beta_g \\ V_\xi(y_i) = \sigma_g^2 \upsilon_i \end{cases}, \quad g = 1,\ldots,G.$$

4.3. *Estimadores indirectos*

Tratam-se de estimadores que utilizam informação relativa à variável de estudo e às variáveis auxiliares fora do domínio de interesse, sem qualquer referência às suas propriedades de não enviesamento no desenho.

A estimação sintética baseia-se no pressuposto de que o domínio é de alguma forma semelhante a outro domínio de maior dimensão ou conjunto de domínios que contenham o primeiro.

O estimador sintético tem normalmente pequena variância, embora possa ser altamente enviesado se o pressuposto estiver errado.

Um estimador sintético simples vem

$$\hat{\tau}_{d,syn,m} = N_d \frac{\sum_{i\in s} \pi_i^{-1} y_i}{\sum_{i\in s} \pi_i^{-1}}, \qquad \textbf{eq. 4.11}$$

que parte do pressuposto que a média da variável de interesse no domínio é igual à média no conjunto da população.

Um estimador sintético menos simplista e de utilização mais comum é baseado na estratificação ou pós-estratificação

$$\hat{\tau}_{d,sym,st,m} = \sum_h N_{h,d} \frac{\sum_{i \in s_h} \pi_i^{-1} y_i}{\sum_{i \in s_h} \pi_i^{-1}}. \qquad \textbf{eq. 4.12}$$

Aqui a hipótese de base é menos restritiva. Supõe-se que em cada estrato ou pós-estrato a média da variável de interesse é constante para todos os domínios. Se de facto o critério de estratificação for de boa qualidade, esta hipótese poderá ser mais aderente à realidade pois a variável de interesse em cada estrato tenderá a apresentar um comportamento homogéneo.

Os estimadores sintéticos poderão igualmente recorrer a informação auxiliar para além das dimensões das sub-populações.

Podem assim propôr-se vários estimadores sintéticos por rácio

$$\hat{\tau}_{d,syn,r} = \tau_{xd} \hat{R} = \tau_{xd} \frac{\hat{\tau}_\pi}{\hat{\tau}_{x,\pi}}, \qquad \textbf{eq. 4.13}$$

$$\hat{\tau}_{d,syn,st,r} = \sum_h \tau_{xh,d} \hat{R}_h = \sum_h \tau_{xh,d} \frac{\hat{\tau}_{h,\pi}}{\hat{\tau}_{xh,\pi}}, \qquad \textbf{eq. 4.14}$$

ou ainda uma formulação alternativa proposta por Singh e Tessier (1976) [4]

$$\tilde{\tau}_{d,syn,r} = \tau_{xd} \frac{\hat{\tau}_\pi}{\tau_x}. \qquad \textbf{eq. 4.15}$$

Estes estimadores têm sido estudados por Gonzalez (1973), Gonzalez e Waksberg (1973), Ghangurde e Singh (1977,1978), entre outros.

Situações há ainda, onde é possível dispôr de informação relativa a uma segunda variável auxiliar z.

---

[4] Tanto $\hat{\tau}_{d,sysn,r}$ como $\tilde{\tau}_{d,sysn,r}$ apresentam o mesmo enviesamento sintético. A escolha entre os dois estimadores depende normalmente da correlação entre $\hat{\tau}_\pi$ e $\hat{\tau}_{x,\pi}$. É possível demonstrar que para grandes amostras $V(\hat{\tau}_{d,sysn,r}) \leq V(\tilde{\tau}_{d,sysn,r})$ se $\rho \geq 0{,}5 C_x / C_y$ onde C_y e C_x são os coeficientes de variação de $\hat{\tau}_\pi$ e $\hat{\tau}_{x,\pi}$ respectivamente. Em alguns casos, quando a correlação entre os dois estimadores é moderada e o coeficiente de variação de x grande, $\tilde{\tau}_{d,sysn,r}$ será preferível.

Neste caso, um estimador sintético por rácio bivariado pode vir na forma

$$\hat{\tau}_{d,syn,r2} = \gamma_d \tau_{xd} \frac{\hat{\tau}_\pi}{\hat{\tau}_{x,\pi}} + (1-\gamma_d)\tau_{xd} \frac{\hat{\tau}_\pi}{\hat{\tau}_{z,\pi}}, \qquad \textbf{eq. 4.16}$$

onde γ_d é uma escolha adequada.

Finalmente apresenta-se um estimador sintético mais genérico baseado no estimador pela regressão

$$\hat{\tau}_{d,sysn,reg} = \tau_{xd}\hat{B}, \qquad \textbf{eq. 4.17}$$

com

$$\hat{B} = \left[\sum_{i\in s} \upsilon_i^{-1}\pi_i^{-1}x_i x_i'\right]^{-1} \sum_{i\in s} \upsilon_i^{-1}\pi_i^{-1}x_i y_i \,. \qquad \textbf{eq. 4.18}$$

Repare-se que está subjacente um modelo do tipo

$$\begin{cases} E_\xi(y_i) = x_i'\beta \\ V_\xi(y_i) = \sigma^2\upsilon_i \end{cases}$$

Os restantes estimadores sintéticos apresentados podem mais uma vez ser encarados como casos particulares do estimador sintético pela regressão.

Uma variante do estimador eq. 4.17 sugerida por Royal (1979), tem a forma

$$\hat{\tau}_{d,sysn,Roy} = \sum_{i\in s_d} y_i + \left(\tau_{xd} - \sum_{i\in s_d} x_i\right)\hat{B}, \qquad \textbf{eq. 4.19}$$

onde apenas são estimados sinteticamente os valores da variável de interesse fora da amostra.

Torna-se agora claro que os estimadores directos modificados podem ser encarados como estimadores sintéticos com um ajustamento *design based* para o enviesamento.

De facto, repare-se que o estimador directo modificado pela regressão eq. 4.9 pode ser escrito na forma

$$\begin{aligned} \hat{\tau}_{d,sreg} &= \tau_{xd}\hat{B} + \left(\hat{\tau}_{.d} - \hat{\tau}_{xd}\hat{B}\right) \\ &= \hat{\tau}_{d,syn,reg} + \sum_{i\in s_d} \pi_i^{-1} e_i \end{aligned} \qquad \textbf{eq. 4.20}$$

onde os e_i representam os resíduos da regressão (diferença entre os valores reais e valores previstos da variável de interesse).

A quantidade $\sum_{i \in s_d} \pi_i^{-1} e_i$ é um estimador do erro total cometido na previsão com base no estimador sintético e pode ser encarada como um termo de ajustamento para o enviesamento, i.e. um estimador do enviesamento do estimador sintético.

Quando a dimensão da amostra no domínio é muito pequena, a utilização deste termo de correcção aumenta substancialmente a variância do estimador, podendo em certas situações ser preferível (de acordo com as expectativas existentes) recorrer a um estimador sintético necessariamente enviesado, mas de menor variância.

Usar um estimador sintético é de alguma forma um jogo. Este poderá ser atractivo no caso de se verificar a expectativa de um pequeno enviesamento; no entanto está sempre presente um risco de não validade dos pressupostos que poderá conduzir à construção de intervalos de confiança inválidos.

Na verdade, não se deverá ignorar que um estimador sintético poderá ter um enviesamento que contribua significativamente para o seu EQM. Como tal, mesmo em situações em que o EQM seja pequeno (mesmo que menor do que a variância de um estimador *design based* para o mesmo parâmetro), a eventual existência de um forte rácio de enviesamento poderá levar à invalidade dos intervalos de confiança que venham a ser construídos com base na estimativa pontual e no EQM do estimador.

Em qualquer caso, a estimação indirecta só deverá ter lugar quando não for possível obter estimativas *design based* com precisão adequada, mesmo depois da utilização da informação auxiliar.

Embora não seja manifestamente objectivo deste artigo desenvolver medidas de precisão dos estimadores apresentados, nem métodos para a sua estimação, será útil apresentar os EQM teóricos aproximados para alguns dos estimadores analisados, a fim de clarificar os seus méritos relativos.

Para o que se segue, considera-se o enquadramento de uma sondagem aleatória simples realizada sobre uma população com D domínios, tendo sido estimado o total da variável de interesse para um domínio d através dos seguintes estimadores: estimador directo pela regressão (1), estimador directo modificado pela regressão (2), estimador sintético pela regressão (3).

Tem-se então para os erros quadráticos médios teóricos dos 3 estimadores,

$$EQM(1|n_d) = F(n_d^{-1}(1+n_d^{-1})) + O(n_d^{-2})$$
$$EQM(2|n_d) = F(n_d^{-1}(1+n^{-1})) + O(n^{-2})$$
$$EQM(3|n_d) = F(n^{-1}(1+n^{-1})) + O(n^{-2}) + K$$

Para o estimador directo (1) tem-se que quer a variância, quer o enviesamento, são de ordem n_d^{-1}, apresentando ainda uma contribuição de 2ª ordem para a variância de ordem n_d^{-2}.

No estimador modificado (2) recorde-se que é utilizada a totalidade da amostra para estimar os coeficientes da regressão. Aqui a variância básica permanece de ordem n_d^{-1} e será na generalidade dos casos práticos maior do que a variância básica do estimador (1). A vantagem está essencialmente na redução do enviesamento que vem de ordem n^{-1} e da contribuição de 2ª ordem para a variância que é agora de ordem $n_d^{-1}n^{-1}$.

Finalmente para o estimador sintético (3), tem-se uma variância de ordem n^{-1}. Desta forma, a redução da variância deste estimador, quando comparado com os outros dois, poderá ser substancial. O preço a pagar por esta redução de variância consiste no enviesamento que é agora de ordem 1. Este é de facto independente da dimensão da amostra e só seria nulo no caso limite (improvável na prática) em que os parâmetros da regressão fossem exactamente os mesmos para o domínio e para o conjunto da população.

4.4. *Estimadores combinados*

Um estimador combinado incorpora uma combinação entre uma componente *design based* e uma componente *model based*.

Tipicamente apresenta a forma de uma média ponderada de um estimador *design based* e de um estimador sintético

$$\hat{\tau}_{d,com} = \lambda_d \hat{\tau}_{d,des} + (1-\lambda_d)\hat{\tau}_{d,syn}. \qquad \textbf{eq. 4.21}$$

A ideia subjacente é a de equilibrar o enviesamento potencial do estimador sintético com a instabilidade do estimador *design based*.

Estes estimadores poderão ainda ser classificados segundo 3 grandes tipos, de acordo com as possíveis abordagens para a definição dos pesos λ_d: pesos fixados à partida, dependentes da dimensão da amostra e dependentes dos dados.

A abordagem mais simples consiste em fixar os pesos à partida de acordo com as expectativas do analista.

Tal tem a desvantagem óbvia de não ter em conta (pelo menos objectivamente) a informação disponível e em particular a fiabilidade observada do estimador por desenho.

De facto, para algumas amostras realizadas o estimador por desenho será mais fiável do que para outras (não se esqueça que a dimensão da amostra no domínio é uma variável aleatória), pelo que terá sentido que o peso reflicta tal facto.

No caso em que o estimador é dependente da dimensão da amostra, tomam-se os pesos como funções do quociente $\hat{N}_d / N_d$.

Drew, Singh e Choudhry (1982) propõem um estimador dependente da dimensão da amostra

$$\hat{\tau}_{d,ssd,r} = \lambda_d \hat{\tau}_{d,r} + (1-\lambda_d)\hat{\tau}_{d,syn,r}, \qquad \textbf{eq. 4.22}$$

onde

$$\lambda_d = \begin{cases} 1 & se \quad \hat{N}_d \geq \delta N_d \\ \hat{N}_d / \delta N_d & \text{cc} \end{cases}$$

e δ é escolhido subjectivamente para controlar a contribuição da componente sintética.
A ideia é fácil de compreender mesmo intuitivamente e torna-se mais evidente quando se enquadra o problema no âmbito de uma sondagem aleatória simples. Neste caso, para $\delta = 1$, o peso λ_d vem igual a 1, reduzindo-se consequentemente o estimador combinado ao estimador *design based*, quando $n_d \geq E(n_d)$, onde n_d é a dimensão efectiva da amostra no domínio.

De facto, será natural que o peso atribuído à componente *design based* seja tanto maior quanto maior for a dimensão efectiva da amostra. Será atribuído um peso considerável à componente sintética quando a dimensão da amostra no domínio é muito pequena. Este peso é gradualmente transferido para o estimador *design based* à medida que a referida dimensão amostral cresce, de tal forma que o estimador combinado seja consistente. Implicitamente está-se a considerar que o valor esperado de n_d corresponde a uma dimensão amostral suficiente para produzir estimativas directas com um grau de precisão aceitável. Este raciocínio terá obviamente especial sentido quando se trabalha com domínios planeados, caso em que as amostras poderão ser desenhadas de forma a que as amostras dos domínios tenham dimensão com valor esperado adequado.

Outro estimador de natureza semelhante é sugerido por Särndal e Hidiroglou (1989)

$$\hat{\tau}_{d,ssd,reg} = \lambda_d \hat{\tau}_{d,sreg} + (1-\lambda_d)\hat{\tau}_{d,syn,reg}, \qquad \textbf{eq. 4.23}$$

onde

$$\lambda_d = \begin{cases} 1 & se \quad \hat{N}_d \geq N_d \\ \left(\hat{N}_d / N_d\right)^{h-1} & \text{cc} \end{cases}$$

Uma das principais preocupações que deverá estar associada à definição dos pesos é a de assegurar que o enviesamento da componente sintética do estimador seja mantido dentro de limites aceitáveis.

O estimador eq. 4.22, adaptado ao caso da regressão, tem tido grande utilização prática, sendo habitual definir o parâmetro δ no intervalo $[2/3;3/2]$, dependendo obviamente do risco de enviesamento que se pretenda correr.

Para o caso do estimador combinado dependente dos dados, os pesos óptimos para a combinação das duas componentes depende normalmente do EQM de cada uma delas e da sua covariância.

Estas quantidades são habitualmente desconhecidas, podendo ser derivadas teoricamente com relativa facilidade. No entanto, a sua estimação com alguma fiabilidade torna-se mais complexa, sendo fácil de conceber que no caso em que a dimensão das amostras nos domínios é insuficiente para produzir estimativas directas, também não será adequada para produzir estimativas das variâncias e enviesamentos correspondentes.

Uma via possível para este fim passa pela modelização do enviesamento da parte sintética, conduzindo à construção de estimativas indirectas para estes parâmetros.

Uma abordagem bem conhecida, que tem constituido de alguma forma um paradigma para muitas das abordagens subsequentes, é devida a Fay e Herriot (1979).

Nesta abordagem, os enviesamentos dos estimadores sintéticos para os domínios são modelizados como efeitos aleatórios independentes com variância desconhecida mas fixa, o que acaba por resultar num estimador combinado com um enviesamento consideravelmente reduzido quando comparado com um estimador puramente sintético.

No fundo, esta pode ser considerada como uma abordagem Bayesiana empírica baseada no melhor previsor linear centrado (BLUP).

O estimador de Fay-Herriot aparece então na forma de uma combinação convexa de um estimador directo e de um estimador sintético.

Suponha-se que os totais da variável de interesse para cada um dos D domínios da população estão ligados a um vector de variáveis explicativas, através de um modelo linear

$$\tau_d = \tau_{xd}\beta + a_d\,, \qquad \textbf{eq. 4.24}$$

onde os a_d's representam efeitos aleatórios associados aos domínios, não correlacionados, com média nula e variância w_d que se assume conhecida.

Virá em notação matricial

$$\tau = \tau_x \beta + a . \qquad \textbf{eq. 4.25}$$

O modelo aparece assim como a soma de uma componente estrutural, $\tau_x \beta$, com uma componente de domínio, a.

Considera-se uma modelização dos estimadores directos para os domínios

$$\hat{\tau}_{dir} = \tau_x \beta + a + \varepsilon , \qquad \textbf{eq. 4.26}$$

onde $\hat{\tau}_{dir}$ é o vector dos estimadores directos para os D domínios e os ε_d são os erros da sondagem, não correlacionados, com média nula e variância v_d, $\varepsilon \cap (0, V)$. Assume-se igualmente que a é não correlacionado com ε.

O BLUP de τ sob o modelo é

$$\begin{aligned} \hat{\tau}_{FH} &= \hat{\tau}_{syn} + \Lambda\left(\hat{\tau}_{dir} - \hat{\tau}_{syn}\right) \\ &= \Lambda \hat{\tau}_{dir} + (1 - \Lambda)\hat{\tau}_{syn} \end{aligned} \qquad \textbf{eq. 4.27}$$

onde

$$\Lambda = \left(V^{-1} + W^{-1}\right)^{-1} V^{-1}$$

$$V = \mathrm{diag}(v_1, \ldots, v_D)$$
$$W = \mathrm{diag}(w_1, \ldots, w_D)$$

e a componente sintética $\hat{\tau}_{syn} = \tau_x \hat{B}$, com $\hat{B}$ a estimativa dos mínimos quadrados ponderados (WLS) de β sob o modelo.

Para cada domínio d virá

$$\hat{\tau}_{d,FH} = \frac{w_d}{w_d + v_d} \hat{\tau}_{d,dir} + \frac{v_d}{w_d + v_d} \hat{\tau}_{d,syn} . \qquad \textbf{eq. 4.28}$$

A ideia subjacente parte do modelo especificado para o vector dos totais da variável de interesse nos domínios que compreende duas componentes: $\tau_x\beta$ e a, cujos BLUP são respectivamente $\hat{\tau}_{syn}$ e $\Lambda\left(\hat{\tau}_{dir} - \hat{\tau}_{syn}\right)$.

Curioso será notar que a estrutura do BLUP é a mesma, independentemente de β ser conhecido ou estimado.

Assumiu-se até agora que as matrizes de covariâncias V e W são conhecidas no cálculo do BLUP. Quando estas são substituídas por estimativas, como é habitual na prática, o estimador resultante é designado por EBLUP (Empirical Best Linear Unbiased Predictor).

Quando os ν_d são pequenos, i.e. quando as dimensões efectivas das amostras nos domínios são grandes, o estimador de Fay-Herriot aproxima-se do estimador *design based*, mantendo-se no entanto enviesado condicionalmente a τ, com o enviesamento a tender para zero à medida que os ν_d se tornam pequenos.

Uma protecção contra uma má especificação do modelo poderá ser alcançada através da truncagem da estimativa resultante, no caso de esta se afastar da estimativa directa mais do que uma quantidade múltipla do desvio padrão do estimador directo.

Outras abordagens semelhantes têm sido definidas por diversos autores, em particular Schaible (1979), Battese e Fuller (1981), Drew, Singh e Choudry(82), Battese, Harter e Fuller (1988), Prasad e Rao (1990).

Note-se que a escolha entre uma abordagem dependente da dimensão da amostra ou dependente dos dados poderá ser condicionada por diversos factores. De facto, não se poderá ignorar que a abordagem dependente da dimensão da amostra não reflecte a informação dada pelos valores da variável de interesse, não tendo em conta (directamente) quer a precisão do estimador *design based* quer o possível enviesamento do estimador sintético. No entanto, existem igualmente vantagens óbvias que resultam da maior simplicidade deste estimador e em particular da possibilidade de utilização dos mesmos pesos para a estimação dos totais de todas as variáveis de interesse, criando uma consistência interna que se torna manifestamente útil em situações (como é habitual na prática) em que esteja em causa a estimação de um grande numero de variáveis de interesse[5].

[5] Na verdade, é possível conceber estimadores combinados baseados nos dados, onde a consistência interna seja igualmente assegurada por via de procedimentos multivariados. Tais procedimentos não cabem no âmbito da nossa análise mas poderão ser encontrados em Fuller e Harter (1987), Fay (1987).

5. Métodos de estimação baseados em séries cronológicas

Recentemente têm surgido diversos métodos especiais para estimação em pequenos domínios. Embora a sua análise não constitua objectivo deste artigo, será apresentada pelo seu especial interesse uma generalização do modelo de Fay-Herriot que incorpora informação histórica no processo de estimação através do recurso a modelos baseados em séries cronológicas.

De facto, uma das ideias fundamentais das várias abordagens apresentadas é o recurso a informação auxiliar fora do domínio de estudo.

Quando existem dados históricos disponíveis, nomeadamente no âmbito de uma sondagem repetida no tempo, poderá igualmente ser vantajosa a sua incorporação no processo de estimação. Tal tem a vantagem de tornar os pesos mais estáveis do que se fossem exclusivamente calculados a partir dos dados da sondagem, mas poderá obviamente comprometer a comparação das estimativas referentes a um domínio em vários períodos de tempo.

A ideia base consiste na modelização da relação entre os parâmetros do modelo ao longo do tempo, utilizando então esse modelo para melhorar a eficiência da estimação no período de interesse.

Muitas destas abordagens passam pela generalização do modelo de Fay-Herriot, onde se permite que os parâmetros da regressão e os efeitos dos domínios evoluam de acordo com determinados modelos cronológicos.

Suponha-se então que está disponível informação na forma de estimadores directos para os domínios, em diversos períodos de tempo $t = 1, \ldots, T$.

Define-se $\hat{\tau}_{dir,t}$ o vector das estimativas directas $\hat{\tau}_{d,dir}$ baseadas nos dados do período t.

Considera-se igualmente que os totais do vector de variáveis auxiliares estão disponíveis para os domínios.

Recorde-se que o modelo de Fay-Herriot utiliza informação dos vários domínios para o momento corrente t e é dado pela soma de duas componentes, um previsor (estimador) da componente estrutural do modelo subjacente e um previsor para o efeito aleatório de domínio .

$$\hat{\tau}_{FH,t} = \tau_{x,t}\hat{B}_t + \hat{a}_t . \qquad \textbf{eq. 5.1}$$

Estando disponível informação sobre vários períodos de tempo, será igualmente possível recorrer a essa informação para efeitos de estimação.

Especifica-se então um novo modelo estrutural para as estimativas directas baseado em séries cronológicas.

Para simplificação de notação, no que se segue define-se

$$\alpha_t = \left(\beta_t', a_t'\right)'$$
$$H_t = \left(\tau_{x,t}, I\right).$$

Tem-se então

$$\hat{\tau}_{dir,t} = \tau_t + \varepsilon_t, \qquad \textbf{eq. 5.2}$$

$$\tau_t = \tau_{x,t}\beta_t + a_t \equiv H_t\alpha_t. \qquad \textbf{eq. 5.3}$$

Permite-se então que os parâmetros da regressão, bem como os efeitos aleatórios dos domínios sintetizados em α_t evoluam no tempo de acordo com o modelo

$$\alpha_t = G_t\alpha_{t-1} + \varsigma_t, \qquad \textbf{eq. 5.4}$$

onde

$$G_t = \begin{bmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{bmatrix}, \quad \varsigma_t = \begin{bmatrix} \xi_t \\ \eta_t \end{bmatrix}.$$

Assume-se ainda que os ε_t e ς_t são não correlacionados, ς_t é não correlacionado com os α_s para $s < t$ e que $\varepsilon_t \cap \left(0, V_t\right)$, $\varsigma_t \cap \left(0, \Gamma_t\right)$,

onde $\Gamma_t = \text{diag bloco}\left\{C_t, Q_t\right\}$.

Assume-se habitualmente que as matrizes V_t, C_t e Q_t são diagonais. Um dos pressupostos implicitos é o de que os erros da sondagem são não correlacionados quer seccionalmente, quer cronologicamente.

É ainda habitual para efeitos práticos assumir que G_t é uma matriz identidade, fazendo com que β_t e a_t evoluam de acordo com um processo do tipo passeio aleatório.

Neste caso particular tem-se que

$$\beta_{jt} = \beta_{jt-1} + \xi_{jt} \quad j = 1,\ldots,J \qquad \textbf{eq. 5.5}$$

e

$$a_{dt} = a_{dt-1} + \eta_{dt} \quad d = 1,\ldots,D, \qquad \textbf{eq. 5.6}$$

onde J é o número de variáveis explicativas do modelo e D o número de domínios, e com $\xi_t \cap (0, C_t)$, $\eta_t \cap (0, Q_t)$.

Começa-se por encontrar $\tilde{\alpha}_T$ o BLUP de α_T, a partir do qual se obtém o estimador óptimo (BLUP) de τ_T baseado em todas as estimativas até ao momento T, como $H_T \tilde{\alpha}_T$.

A partir da teoria dos modelos lineares com efeitos aleatórios torna-se possível obter $\tilde{\alpha}_T$ directamente a partir do conjunto dos dados completos.

No entanto, uma abordagem mais fácil resulta de se reparar que os α_t estão ligados ao longo do tempo através da equação de transição eq. 5.4, pelo que se torna mais conveniente o seu cálculo recursivo utilizando um filtro de Kalman.

Em última análise, um filtro de Kalman é uma técnica Baysiana, no sentido em que em cada momento t, a distribuição posterior de α_t considerando os dados até $t-1$, é actualizada de forma a obter a distribuição posterior de α_t considerando os dados até t.

Claro que não se torna necessário seguir esta perspectiva em situações em que se trabalha sob modelos lineares mistos.

De facto, representando por $\tilde{\alpha}_{T|s}$ o BLUP de α_T baseado nos dados até s, $s < T$ e considerando o modelo acima apresentado, é bem sabido que o BLUP $\tilde{\alpha}_T$ baseado no conjunto dos dados até T, é o mesmo que o BLUP de α_T baseado no $\tilde{\alpha}_{T|s}$ e nas últimas $T-s$ observações.

Por outras palavras, a informação dos dados de períodos de tempo anteriores pode em qualquer momento ser condensada num BLUP apropriado.

Em termos práticos, torna-se necessário estimar os parâmetros do modelo até agora considerados conhecidos (valores das matrizes C_t e Q_t), o que sob o modelo apresentado pode ser conseguido de forma relativamente simples pelo método dos momentos e portanto sem se tornar necessário assumir qualquer tipo de distribuição subjacente. O estimador resultante passa então a ser designado por EBLUP.

Para obter o EBLUP $\tilde{\alpha}_T$ corre-se o filtro de Kalman com os valores iniciais de $\tilde{\alpha}_t$ e seu EQM obtidos a partir do estimador de Fay-Herriot para $t=1$.

Referências a este modelo podem ser encontradas em Singh, Mantel e Thomas (1994). Outras abordagens baseadas em séries cronológicas podem ainda ser encontradas em Pfeffermann e Burck (1990), Pfeffermann (1991), Rao e Yu (1992).

6. Considerações finais

Os ganhos conseguidos através das diversas técnicas para estimação em domínios resultam invariavelmente da utilização de informação auxiliar.

A ideia básica de qualquer método de estimação para pequenos domínios consiste em utilizar informação relativa a outros domínios, assumindo que estes estão ligados por via de um modelo que contem variáveis auxiliares.

Uma das primeiras tarefas do estaticista será a procura de variáveis auxiliares que estejam correlacionadas com a variável ou variáveis de interesse. Em última análise não se deverá esquecer que quanto maior for essa correlação, menor será a variabilidade restante para repartir entre variância de amostragem e variância inter-domínio.

Sempre que possível, é desejável a utilização de estimadores directos que garantam propriedades estatísticas interessantes do ponto de vista do desenho.

Em situações em que não seja manifestamente possível obter estimativas *design based* de precisão aceitável, mesmo depois de se recorrer à utilização de informação auxiliar, nomeadamente em domínios de muito pequena dimensão, torna-se necessário recorrer a outros métodos de estimação.
Os estimadores sintéticos constituem uma opção possível devido à sua relativa facilidade de utilização, apresentando normalmente pequenas variâncias resultantes do agrupamento de dados referentes a vários domínios. A sua grande desvantagem reside no aparecimento de enviesamentos de valor desconhecido, que poderão ser frequentemente apreciáveis, pondo assim em causa a validade de eventuais intervalos de confiança que venham a ser construídos com base nestes estimadores.

Outra opção consiste no recurso a estimadores combinados. Estes estimadores resultam habitualmente da combinação de uma componente *design based* com uma componente sintética. Desta forma, torna-se possível obter estimadores consistentes que apresentam normalmente enviesamentos menores do que os associados aos estimadores exclusivamente sintéticos.

Neste contexto têm vindo a surgir abordagens baseadas na utilização de dados históricos, juntamente com os dados seccionais.

Tem sido verificado empiricamente, nomeadamente através de simulação de Monte Carlo que os modelos baseados em séries cronológicas apresentam melhor performance do que os baseados simplesmente em dados seccionais no que diz respeito a enviesamento e EQM.

Tal facto não é aliás de estranhar. O principal problema na estimação em pequenos domínios reside na escassez de dados disponíveis, resultantes da pequena dimensão das amostras, o que em última análise condiciona a qualidade dos resultados obtidos, apesar de todos os esforços que possam ser efectuados no desenvolvimento de modelos aderentes à realidade.

É sobretudo no âmbito de amostragem repetida no tempo que poderão surgir oportunidades de desenvolvimento de abordagens robustas, já que neste contexto a quantidade de informação disponível pode aumentar substancialmente.

Não se deverá igualmente ignorar que independentemente da abordagem seguida, na prática deverão sempre ser levados a cabo diagnósticos adequados para os dados da sondagem, antes de se optar por determinado método de estimação assistido por modelo.

Bibliografia

Battese, G.E., e Fuller, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

Battese, G.E., Harter, R.M., e Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Cassel, C.M., Särndal, C.E., eWretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Drew, J.D., Singh, M.P. e Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.

Fay, R.E. (1987). Application of multivariate regression to small domains estimation. In *Small Area statistics*. (Eds. R. Platek et al.). New York: Wiley.

Fay, R.E., e Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fuller, W.A., e Harter, R.M. (1987). The multivariate components of variance model for small area estimation. *In Small Area statistics*. (Eds. R. Platek et al.). New York: Wiley.

Ghangurde, O.D., e Singh, M.P. (1977). Synthetic estimates in periodic household surveys. *Survey Methodology*, 3, 152-181.

Ghangurde, O.D., e Singh, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.

Ghosh, M., e Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9, 55-93.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

Gonzalez M.E., e Waksberg, J. (1973). Estimation of the error of synthetic estimates. Apresentado na primeira reunião da International Association of Survey Statisticians, Vienna, Austria.

Isaki, C.T., e Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.

Pfeffermann, D., e Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

Prasad, N.G.N., e Rao, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K., e Yu, M. (1992). Small Area Estimation by combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

Royal, R.M. (1979). Prediction models in small area estimation. In *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Catálogo da Library of Congress nº 79-600067, 36-53.

Särndal, C.E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

Särndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

Särndal, C.E., e Hidiroglou, M.A. (1989). Small Domain Estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

Särndal, C.E., Swensson, B., e Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, A.C., Mantel, H.J. e Thomas B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

Singh, M.P., Gambino, J., e Mantel, H.J. (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, 20, 3-22.

Singh, M.P., e Tessier, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

Schaible, W.L. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.