Inspecções de software usando modelos de captura-recaptura na estimação de falhas

Russell Alpizar-Jara
Universidade de Évora, DMAT e CIMA - alpizar@uevora.pt

Paulo Infante
Universidade de Évora, DMAT e CIMA - pinfante@uevora.pt

André Martins
Universidade de Évora, DMAT e CIMA - jimimartins@gmail.com

Resumo: Nas últimas décadas, os modelos de captura-recaptura em populações fechadas têm sido muito utilizados para estimar o número de falhas num sistema. Modelos que tomam em conta a heterogeneidade das probabilidades de detecção de falhas, diferenças nas probabilidades de detecção por parte dos avaliadores, e simplificações destes modelos têm sido propostos na literatura (ex. M_{th}, M_h, M_t e M_0). Uma classe específica de modelos que tem sido amplamente negligenciada, considera diferenças entre a probabilidade da primeira detecção de uma dada falha e as probabilidades de detecção subsequentes. Poderíamos pensar que não há razão para suspeitar, que após a primeira detecção duma falha por parte de um avaliador, a probabilidade de detecção da mesma falha seja diferente para os outros avaliadores. Este tipo de heterogeneidade é conhecido na literatura inglesa sobre vida selvagem por "behavioural response", podendo ser associada a uma potencial falta de independência entre os avaliadores. Admita-se, por exemplo, que após a falha ser detectada pela primeira vez por um avaliador, os restantes avaliadores são alertados de alguma forma para o tipo de falha. Neste estudo, repetindo a análise de um conjunto de dados relativos a uma experiência controlada apresentada em Thelin *et al.* (2002), observamos que são também possíveis os modelos do tipo M_b, M_{bh}, e M_{tb}. Discutimos sobre a validade da hipótese de independência entre os avaliadores e propomos que os modelos do tipo M_b possam ser vistos como um indicador de uma potencial violação deste pressuposto. Analisamos alternativas para testar esta hipótese e outros aspectos relacionados com a aparente independência do número de avaliadores.

Palavras–chave: captura-recaptura, inspecção de software, fiabilidade do software

Abstract: During the last decades, closed population capture-recapture models have been widely used to estimate the number of faults in a system. Models that jointly account for heterogeneity in faults detection probabilities, differences in reviewers' detection probabilities, and simplifications of these models have been proposed in the literature (i.e M_{th}, M_h, M_t and M_0). A special class of models, that has been largely neglected, is one that account for differences between first and subsequent fault detection probabilities. One may think that there is no reason to believe that after detecting a given fault by a reviewer the first time, its detection probability would be different

for other reviewers. This sort of heterogeneity, known as "behavioural heterogeneity" in the wildlife literature, could be associated to a potential lack of independence among reviewers. Suppose for instance, that after first detection of a fault by any inspector, other reviewers became some how aware of the type of fault. In this study, we re-analyzed a data set from a controlled experimental setting by Thelin *et al.* (2002), and noticed that models of the type M_b, M_{bh}, e M_{tb} were also possible. We question the assumption of independent detection among reviewers and proposed that M_b-type models could be used as an indicator of potential violation of this assumption. We tried out different approaches for data analyses to assess this hypothesis and other issues related to the number of seemingly independent reviewers.

Keywords: capture-recapture, quality control, software reliability

1 Introdução

Os modelos de captura-recaptura têm sido amplamente utilizados para estimar a abundância e parâmetros demográficos em várias populações de interesse. Por exemplo, na área das Ciências Biológicas e Ecologia (pássaros, mamíferos, répteis, peixes, insectos e plantas); nas áreas da Ciências Médicas, Saúde Pública e Epidemiologia (populações humanas evasivas e/ou com doenças crónicas); na área das Ciências Sociais (imigração ilegal e os sem abrigo entre outros). Mais recentemente esta metodologia tem sido muito aplicada na estimação de falhas em software de computadores. Seber (1982) refere a potenciais aplicações em problemas de edição de texto, correcção de programas de computação, e controlo de qualidade.

Os pressupostos básicos dos modelos de populações fechadas são: a população é fechada e permanece constante no período de estudo (não há nascimentos, mortes, emigração ou imigração), as marcas não se perdem nem podem ser ignoradas pelos observadores. Oito modelos têm sido propostos e desenvolvidos para tratar o problema da heterogeneidade nas probabilidades de captura considerando três fontes de variação: t = variação temporal, b = comportamento, h = heterogeneidade individual. Estes modelos podem apresentar-se numa estrutura hierárquica como se ilustra na (Figura 1), onde o modelo mais geral M_{tbh} considera as três fontes de variação em simultâneo e o modelo nulo M_o assume que as probabilidades de captura são constantes. A selecção de modelos é tipicamente efectuada através de um critério baseado numa função discriminante multivariada que considera vários testes de ajustamento e validação dos modelos ajustados. Este procedimento está disponível no software de distribuição livre, programa CAPTURE (Otis *et al.* 1978 e White *et al.* 1982).

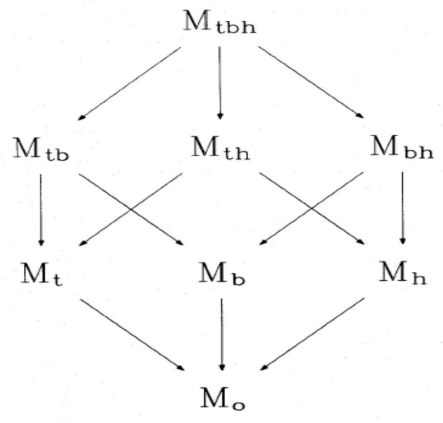

Figura 1: Estrutura hierárquica dos modelos usuais para tratar a heterogeneidade nos problemas de captura (adaptada de White *et al.* 1982).

2 Captura-recaptura em inspecção de software

Os modelos que são actualmente utilizados nas aplicações de inspecção de *software* geralmente tomam em conta a heterogeneidade nas probabilidades de detecção das falhas e as diferenças nas probabilidades de detecção entre inspectores. Simplificações destes modelos também aparecem na literatura (i.e. M_{th}, M_h, M_t and M_o), como pode ver-se, por exemplo, em Peterson *et al.* (2004). Em relação aos modelos do tipo M_b, M_{bh}, M_{tb}, etc., Briand *et al.* (2000), entre outros autores, argumentam que *"In an inspection context, this may be usable to model the fact that defects captured by more than one inspector have usually a higher probability of being detected. However, the estimators for this source of variation depend on the order of trapping occasions (i.e., inspectors). Since no ordering of inspectors seems reasonable in the context of inspections, these estimators are not considered adequate."* Com este trabalho, contestamos estas conjecturas e argumentamos que podem existir razões que justificam a ordenação dos inspectores, segundo as suas habilidades para detectar falhas. Por outro lado, a possível inter-comunicação entre inspectores sobre o tipo de falhas detectadas pode também comprometer estas afirmações, uma vez que situações de "*trap happy*" ou "*trap shy*" podem ocorrer. Na Tabela 1 apresentamos um exemplo hipotético com a história das capturas de 4 indivíduos (falhas) e 5 momentos de amostragem (inspectores). Os dados correspondentes ao indivíduo n^0 2 representam uma situação típica de "*trap shy*" e o n^0

4 representa uma situação de "*trap happy*". Uma situação é "*trap happy*" se após detectada uma falha a primeira vez, os restantes inspectores (numerados a seguir) também detectam a mesma falha. A situação "*trap shy*" seria em princípio menos comum, mas poderá acontecer se a falha não é detectada por algum outro inspector após a primeira detecção.

Tabela 1: Exemplo hipotético história das capturas

Momento de captura (Inspectores)					
indivíduo (falha)	1	2	3	4	5
1	1	1	0	0	0
2	1	0	0	0	0
3	1	0	0	1	0
4	0	1	1	1	1

2.1 Verosimilhança e estimação de parâmetros

Nos modelos para populações fechadas, a estimação de parâmetros é tipicamente feita com recurso ao método de estimação por máxima verosimilhança e também com alguns métodos não paramétricos. Várias abordagens têm sido propostas na literatura sobre vida selvagem e, como indicado anteriormente, algumas dessas abordagens têm sido aplicadas no contexto de sistemas de inspecção de software.

Uma formulação geral da função de máxima verosimilhança para os modelos em populações fechadas é dada por:

$$L(\cdot) = \prod_{i=1}^{N} \prod_{j=1}^{k} p_{ij}^{x_{ij}} (1 - p_{ij})^{1 - x_{ij}} \tag{1}$$

$$\text{com} \quad x_{ij} = \begin{cases} 1 & \text{se o animal } i \text{ é capturado na amostra } j \\ 0 & \text{outros casos} \end{cases}$$

Sendo p_{ij} a probabilidade de que o animal i seja capturado na amostra (ou tempo de amostragem) j, $i = 1, \ldots, N$, e $j = 1, \ldots, k$, uma descrição simbólica dos modelos que consideram heterogeneidade devida ao comportamento dos indivíduos é a seguinte:

- M_b: $p_{ij} = p$ na 1^a captura, $p_{ij} = c$ para as recapturas.
- M_{bh}: $p_{ij} = p_i$ na 1^a captura, $p_{ij} = c_i$ para as recapturas.
- M_{tb}: $p_{ij} = p_j$ na 1^a captura, $p_{ij} = c_j$ para as recapturas.
- M_{tbh}: $p_{ij} = p_{ij}$ na 1^a captura, $p_{ij} = c_{ij}$ para as recapturas.

Os parâmetros destes modelos são o tamanho da população, N, e as probabilidades $\{p_{ij}\}$. Devido ao elevado número de parâmetros em alguns destes modelos, a função de verosimilhança (1) apresenta problemas de identificabilidade e estimabilidade se não forem consideradas algumas restrições. Devido a estes problemas, autores com Burnham e Overton (1978) e Chao *et al.* (1992) propõem estimadores alternativos baseados em métodos não paramétricos.

No contexto de inspecção de software, N representa o número total de potenciais falhas num texto (parâmetro de interesse que pretendemos estimar), k é o número de inspectores, e as probabilidades p_{ij} representa a probabilidade de que a falha i seja detectada pelo inspector j, $i = 1, \ldots, N$, e $j = 1, \ldots, k$. Certamente que não faz sentido pensar que as falhas "reagem" à eventual detecção dos inspectores. No entanto, na eventualidade dos inspectores poderem comunicar sobre algumas características das falhas detectadas, poderá incidir na possível violação deste pressuposto.

2.2 Exemplo ilustrativo

Usamos dados de um estudo publicado por Thelin *et al.* (2002) acerca de uma experiência controlada conduzida na Suécia. O estudo pretendeu avaliar um sistema de software de gestão de táxis usando 27 estudantes universitários como inspectores. O valor da verdadeira população era conhecido e consistia em 37 falhas. No entanto, os inspectores, no seu conjunto, só conseguiram encontrar 33 falhas após 270 detecções. O objectivo do estudo foi estimar o número total de falhas, baseados nos dados recolhidos. É evidente que algumas das falhas têm uma probabilidade muito baixa ou quase nula de ser detectada. Nas Figuras 2 e 3 apresentamos, respectivamente, gráficos de barras que mostram a distribuição de frequências do número de vezes que uma determinada falha foi detectada e o número de falhas que detectou cada inspector. Tanto as falhas como os inspectores foram codificados sem ordem aparente. Thelin *et al.* (2002) não referem nas suas análises a modelos que envolvem a fonte de variação nas probabilidades de detecção do *"b"*, comportamento, mas se incluirmos estes modelos no procedimento de selecção implementado no software CAPTURE (Otis *et al.* 1978 e White *et al.* 1982), notamos que estes modelos também são possíveis como mostramos a seguir. A nossa estratégia de análises consiste em estabelecer vários cenários para mostrar a utilidade efectiva e o potencial dos modelos do tipo *"b"*, comportamento, quando aplicados na inspecção de sistemas (neste caso em concreto inspecção de falhas em software de computação).

3 Análise e Resultados

Mostraremos agora que no estudo de Thelin *et al.* (2002), os modelos do tipo M_b, M_{bh} e M_{tb} também são possíveis. Examinamos 5 cenários que passamos a denotar com as letras **O, A, B, C** e **D** respectivamente.

O cenário **"O"** refere-se a uma análise dos dados "originais", tal e como foram publicados no artigo de Thelin *et al.* (2002). A primeira variante desta análise

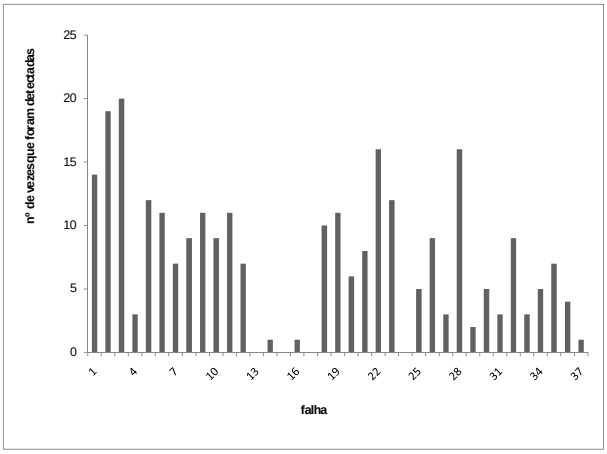

Figura 2: Distribuição de frequências com que foram detectadas as falhas

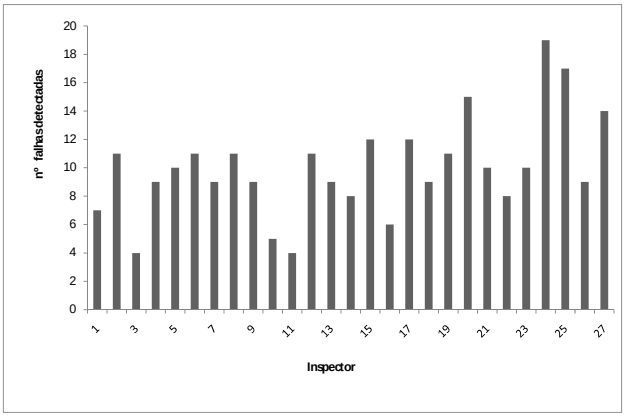

Figura 3: Número de falhas detectadas por cada inspector

"O_1" considera todos os 27 inspectores em nenhuma ordem particular. Posteriormente dividimos os dados em dois subconjuntos para reduzir o número de inspectores mais ou menos a metade. Na segunda variante, "O_2", optamos por

analisar os dados dos primeiros 13 inspectores e na terceira variante, "O_3", analisamos os dados dos 14 inspectores restantes. Salientamos que neste primeiro cenário não foi considerada qualquer ordem, excepto a disposição dos dados tal como foram publicados.

No cenário **"A"** introduzimos um certo grau de ordem para apreciar o efeito confundido de heterogeneidade dos tipos "b" e "h". Ordenamos os inspectores segundo o número de falhas detectadas por cada um deles e procedemos com a divisão dos dados em 3 subconjuntos para análise. "A_1" representa o subconjunto dos 13 inspectores que detectaram menos falhas (entre 4 e 9 falhas). Denominamos este grupo por **"inspectores maus"**. "A_2" representa o subconjunto dos 14 melhores inspectores que detectaram mais falhas (entre 10 e 19 falhas). Denominamos este grupo por **"inspectores bons"**. Finalmente temos um grupo dos **"inspectores nem bons nem maus"**, "A_3". Este grupo está formado por 14 inspectores que detectaram entre 9 e 11 falhas.

No cenário **"B"** cria-se uma maior hierarquia na ordem para constituir grupos mais homogéneos eliminando algum efeito da ordem dos inspectores, e no grau de heterogeneidade nas probabilidades de detecção. Dividimos agora os dados em 4 grupos, reduzindo ainda mais o número de inspectores por análise. "B_1" representa o subconjunto dos 7 inspectores que detectaram menos falhas (entre 4 e 8 falhas). Denominamos este grupo por **"inspectores piores"**. "B_2" representa o subconjunto de 6 inspectores que detectaram 9 falhas cada. Denominamos este grupo por **"inspectores menos bons"**. "B_3" representa um subconjunto de 8 inspectores que detectaram 10 ou 11 falhas. Denominamos este grupo por **"inspectores bons"**. Finalmente, "B_4" é um grupo de 6 melhores inspectores que detectaram entre 12 e 19 falhas e que denominamos **"inspectores melhores"**.

No cenário **"C"** consideramos novamente todos os dados, mas ordenamos não só segundo o número de falhas detectadas por cada inspector, mas também segundo a ordem em que cada falha foi detectada por um ou mais inspectores. Em "C_1" todos os dados são ordenados de maior a menor considerando as ordens dos inspectores e das falhas. Em "C_2" os dados são ordenados de menor a maior considerando os inspectores e de maior a menor considerando as falhas. Finalmente, em "C_3" todos os dados são ordenados de menor a maior considerando as ordens dos inspectores e das falhas.

No último cenário, **"D"**, reduzimos o número de inspectores a 4 (indivíduos ou equipas). Numa primeira abordagem "D_1" juntamos dados dentro de cada grupo "B_1'' a "B_4", considerando cada grupo uma equipa, e cada equipa como um estrato. Consideramos também a selecção aleatória de um inspector dentro de cada estrato e analisamos os dados dos 4 inspectores em ordem crescente ("D_2") e decrescente ("D_3"), segundo o número de falhas detectadas. Os inspectores seleccionados detectaram respectivamente 4, 9, 11, e 14 falhas.

Na Tabela 2 apresentamos os principais resultados destas análises destacando em particular o modelo seleccionado (a evidência empírica de heterogeneidade do tipo "b"), a estimativa do número total de falhas ($\hat{N}$), o erro padrão (ep), um

Tabela 2: Resultados das análises dos dados em Thelin *et al.* (2002) para os cenários **O, A, B, C** e **D**.

Cenário	Modelo	$\hat{N}$	ep	$95\%IC$	$\%env$	k	n
O_1	M_{tb}	47	22,23	(33;75)	+27,0	27	33
O_2	M_{tbh} *ou* M_h	23	1,46	(23;30)	-37,8	13	22
O_3	M_b *ou* M_h	34	2,97	(34;51)	-8,1	14	33
A_1	M_{bh}	29	1,56	(29;37)	-21,6	13	28
A_2	M_{bh}	34	2,97	(34;51)	-8,1	14	33
A_3	M_{tbh} *ou* M_h	35	6,48	(31;62)	-5,4	14	29
B_1	M_{tbh} *ou* M_0	25	2,73	(23;35)	-32,4	7	22
B_2	M_{bh} *ou* M_{th}	26	2,87	(25;39)	-29,7	6	24
B_3	M_h *ou* M_0	34	5,07	(29;52)	-8,1	8	27
B_4	M_h *ou* M_0	34	3,39	(32;48)	-8,1	6	31
C_1	M_b	33	1,11	(33;40)	-10,8	27	33
C_2	M_b	33	0,002	(33;33)	-10,8	27	33
C_3	M_b	33	0,002	(33;33)	-10,8	27	33
D_1	M_b	33	0,96	(33;39)	-10,8	4*	33
D_2	M_{tbh} *ou* M_0	26	2,89	(22;34)	-29,7	4	22
D_3	M_{th}	30	7,58	(24;60)	-18,9	4	22

*= 4 equipas que envolvem os 27 inspectores

intervalo de confiança associado ($95\%IC$ ou *profile likelihood* quando apropriado), o enviesamento relativo ($\%env = 100 * (\hat{N} - N)/N$), o número de inspectores (k), e as falhas detectadas (n). Uma vez que não temos estimador disponível para M_{tbh} no CAPTURE, quando este modelo é seleccionado apresentamos a estimativa do modelo alternativo. Caso este seja do tipo h, geralmente apresentamos o resultado do estimador proposto por Chao *et al.* (1992).

A presença dos modelos do tipo "b" é evidente. Podemos ainda assinalar que, até certo ponto, é possível controlar o tipo de heterogeneidade induzida nas estimativas, e reduzir ou incrementar o grau de heterogeneidade agrupando os inspectores segundo o nível de experiência nas detecções das falhas.

4 Conclusões e trabalho futuro

Os modelos do tipo M_b, M_{bh} e M_{tb} podem ser utilizados como indicadores para avaliar a eventual violação do pressuposto de independência e heterogeneidade entre inspectores e são ferramentas importantes para avaliar factores de heterogeneidade não observável.

Como trabalho futuro pretendemos avaliar o pressuposto de independência

entre inspectores usando uma abordagem com os modelos log-lineares. Também resulta de interesse avaliar a performance dos estimadores e o método de selecção utilizados em CAPTURE via estudo de simulação.

Apesar de este estudo ser pouco habitual na medida que envolve muitos inspectores e o verdadeiro número de falhas ser conhecido, permite-nos elucidar a importância dos modelos de captura-recaptura que consideram heterogeneidade do tipo b nas probabilidades de detecção de falhas no contexto de inspecção de software. Salientamos que este tipo de modelos tem sido amplamente negligenciado na literatura, quando aplicado neste contexto.

Agradecimentos

Os dois primeiros autores neste trabalho são membros do CIMA-Universidade de Évora, centro de investigação financiado pela Fundação para a Ciência e a Tecnologia (FCT).

Referências

[1] Briand, L., El Emam, K., Freimut, B., e Laitenberger, O. (2000). *IEEE Transactions on Software Engineering*. 26(6), pp. 518-540.

[2] Burnham KP, Overton WS. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65 p. 625-633.

[3] Chao A, Lee SM, Jeng SL. (1992). Estimation of population size for capture-recapture data when capture probabilities vary by time and individual animal. Biometrics 48 p. 201-21.

[4] Otis, D., K. Burnham, G. White, e D. Anderson (1978). Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, (62), pp. 1-135.

[5] Peterson, H., Thelin, T., Runeson, P. e Wohlin, C. (2004). Capture-recapture in software inspections after 10 years research-theory, evaluation and application. *The Journal of Systems and Software* 72, pp. 249-264.

[6] Seber, G.A.F. (1982). *The estimation of animal abundance and related parameters* (2nd ed.), Charles W. Griffin, London.

[7] Thelin, T., Peterson, H. e Runeson, P. (2002). Confidence intervals for capture-recapture estimations in software inspections. *Information and Software Technology* 44, pp. 683-702.

[8] White, G., D. Anderson, K. Burnham, e D. Otis (1982) Capture-Recapture and Removal Methods for Sampling Closed Populations. Technical Report, Los Alamos National Laboratory.