Anexo sobre dados e métodos estatísticos

1. Introdução

1. Os incidentes envolvendo violações de direitos humanos são complexos. Uma testemunha ocular ou vítima pode referir-se a uma ou várias vítimas que podem, cada qual, ter sofrido uma ou várias violações. Cada violação pode também envolver um ou vários perpetradores. Por isso, as interacções entre diferentes pessoas em milhares de incidentes deste tipo requerem o uso cuidadoso de métodos empíricos de identificação e agregação de dados que facilitem análises quantitativas válidas e fiáveis.

2. A Comissão instituiu vários procedimentos para garantir a qualidade dos dados. Neste anexo metodológico são apresentados os dados e métodos que a Comissão utilizou para obter os seus resultados estatísticos.

3. O Anexo está dividido em seis secções principais. Na Secção 1 explica-se a relevância da análise dos dados empíricos para o mandato da Comissão. Na Secção 2 é feita uma descrição detalhada dos diferentes conjuntos de dados que foram usados na análise estatística da Comissão. Na Secção 3 são descritas as técnicas utilizadas de edição e limpeza de dados e de normalização de nomes. Na Secção 4 são apresentadas as tabulações dos registos em diferentes fases do processo de conversão de dados. Na Secção 5 são descritas as diversas técnicas de eliminação de duplicações e de ligação de registos (*record linkage techniques*) usadas para fazer corresponder diversos relatos a uma mesma vítima. Na Secção 6 descreve-se o processamento a que os dados foram submetidos para dar conta de relatos múltiplos de grupos de vítimas anónimas. Finalmente, na Secção 7, são apresentadas as técnicas de estimação estatística que foram usadas para calcular as estimativas totais da magnitude e padrões das violações fatais e deslocações durante o período de referência da Comissão.

A relevância da análise dos dados empíricos para o mandato da Comissão

4. O Grupo de Análise de Dados de Direitos Humanos (*Human Rights Data Analysis Group*, HRDAG) ajudou a Comissão a recolher e analisar os dados sobre violações de direitos humanos relevantes para o período de referência da Comissão, 1974/1999.[*] Neste Anexo explica-se como esses dados foram organizados e processados.

5. A Comissão necessitou dum sistema de gestão de informação para gerir e estruturar os dados necessários, de modo a poder responder às questões colocadas no seu mandato. Em termos mais específicos, o sistema de gestão de informação da Comissão tinha de fornecer informações sobre violações de direitos humanos ocorridas no passado que permitissem subsequentemente:

1. Obter análises estatísticas descritivas dos padrões e tendências genéricos das violações ocorridas, para descrever a "natureza" das violações de direitos humanos (os tipos de violações cometidas).[1]

[*] O HRDAG é uma divisão da Benetech Inc em Palo Alto, Califórnia, USA. O pessoal do HRDAG inclui especialistas em estatística, programação e ligação de registos (*record linkage*). Ao longo dos últimos 20 anos, os membros do HRDAG têm trabalhado em grandes projectos de documentação e análise de violações de direitos humanos nos cinco continentes e em mais de uma dúzia de países. O HRDAG colaborou com comissões oficiais de apuramento da verdade no Haiti, Ãfrica do Sul, Guatemala, Peru, Gana e Serra Leoa; com o Tribunal Penal Internacional para a Antiga Jugoslávia; e com grupos de direitos humanos não-governamentais em El Salvador, Cambodja, Guatemala, Colômbia, Afeganistão, Sri Lanka e Irão. Para mais informações sobre as suas actividades consultar http://www.hrdag.org.

2. Obter projecções estatísticas do número total de violações ocorridas, que permitissem estabelecer a “extensão” das violações de direitos humanos (o número total de violações cometidas).[2]
3. Testar hipóteses estatísticas acerca da regularidade de determinadas violações, para investigar se determinados padrões de violações constituíam "um padrão de abuso sistemático".[3]
4. Efectuar análises de casos através de operações básicas de introdução de registos e pesquisa da base de dados, para descrever os "antecedentes, circunstâncias, factores, contexto, motivos e perspectivas" que originaram violações em grande escala.[4]
5. Obter análises quantitativas estruturadas e testar hipóteses para investigar se "as violações de direitos humanos foram o resultado de planeamento, política ou autorização deliberados” de partes específicas no conflito.[5]
6. Obter explicações formais das metodologias científicas e estatísticas utilizadas, para demonstrar que as conclusões da Comissão se baseiam em "informação factual e objectiva e em dados recolhidos ou recebidos ou postos à sua disposição ".[6]

6. A Comissão estava consciente de que, após terem sofrido violações de direitos humanos, uma parte muito significativa das vítimas e dos seus familiares tinha vivido uma existência de silêncio, medo e isolamento, por vezes durante mais de 25 anos. Por isso, a Comissão teve de conceber sistemas de recolha de dados e de gestão da informação capazes, não apenas de produzirem dados históricos fiáveis, como de promoverem a participação pública no processo de apuramento da verdade.

2. Fontes de dados

7. Esta secção descreve as características as três bases de dados estatísticas que a Comissão criou para levar a cabo uma análise quantitativa das violações de direitos humanos cometidas no passado e para promover a reconciliação em Timor-Leste. A Base de Dados das Violações de Direitos Humanos (*Human Rights Violations Database*, HRVD) é uma colecção de testemunhos narrativos de vítimas, relatórios de tipo qualitativo da Amnistia Internacional (AI) e de dados recolhidos pela Fokupers, uma ONG timorense. O Estudo Retrospectivo de Mortalidade (*Retrospective Mortality Survey*, RMS) é um estudo duma amostra aleatória de agregados familiares que foi usado para medir as deslocações e mortalidade durante o período de referência da Comissão. A Base de Dados do Recenseamento de Cemitérios (*Graveyard Census Database*, GCD) constitui um recenseamento muito completo dos cemitérios públicos nos 13 distritos de Timor-Leste.

8. A combinação dos dados destas três fontes — HRVD, RMS e GCD — permitiu à Comissão elaborar estimativas demográficas independentes sobre a extensão total, padrões, tendências e níveis de responsabilidade pelas violações fatais ocorridas no passado em Timor-Leste.

A Base de Dados das Violações de Direitos Humanos (*Human Rights Violations Database*, HRVD)

9. Nas secções que se seguem são descritos os três projectos de documentação que permitiram constituir a Base de Dados das Violações de Direitos Humanos da Comissão. Também é apresentado o processo de transformação da informação qualitativa desses projectos de documentação em dados estatísticos. Finalmente, apresenta-se também a contabilização dos registos dos três projectos de documentação.

O processo de recolha de testemunhos da Comissão

10. Em Fevereiro de 2003, a Comissão iniciou a recolha de testemunhos narrativos de indivíduos nos 13 distritos de Timor-Leste, bem como de timorenses de leste vivendo em Timor Ocidental. Esses testemunhos constituíram a base da HRVD. A Comissão abriu delegações em cada um dos 13 distritos para executar o seu mandato. Foram recolhidos um total de 7669 testemunhos narrativos relevantes que documentavam violações de direitos humanos. Essas narrativas forneceram informações consideráveis acerca das violações fatais e não fatais durante o período de referência*. O processo de recolha de testemunhos abrangeu todos os 65 subdistritos de cada um dos 13 distritos de Timor-Leste.† Para além da recolha de testemunhos nos distritos, a Comissão também recolheu 86 <s00120> testemunhos de timorenses de leste vivendo em Timor Ocidental, através da sua parceria com uma coligação de ONG sediadas em Timor Ocidental.‡

11. Tendo em conta que os testemunhos foram prestados de forma inteiramente voluntária pelos depoentes e com base numa amostra de conveniência, a distribuição geográfica dos testemunhos não é uniforme. Como se pode ver no gráfico <g5000001>, a comissão recolheu substancialmente mais testemunhos de depoentes em Bobonaro e Ermera do que de depoentes noutros distritos (ver a secção adiante onde se descrevem detalhadamente os possíveis factores que influenciaram o processo de amostragem durante o processo de recolha de testemunhos por parte da Comissão).

[INSERT <g5000001> about here]

12. Para esta informação qualitativa ser analisada estatisticamente, foi codificada numa base de dados FoxPro usando o modelo conceptual para recolha e análise de dados conhecido como *"Who Did What To Whom"* (Quem Fez o Quê a Quem").[7] Embora estes dados forneçam informações muito úteis, o processo de recolha de testemunhos pela Comissão não se baseou numa amostra aleatória de base probabilística. Em vez disso, a Comissão aceitou os testemunhos de todos aqueles que se dispuseram a fornecer as informações de que conseguiam recordar-se. Por esse motivo, os dados de natureza narrativa, isoladamente, não podem ser considerados como estatisticamente representativos da totalidade e padrões de violações em Timor-Leste.

Características demográficas dos depoentes

13. Cerca de 21,4% (1.642/7.669) <s00104> de todos os depoentes no processo de recolha de testemunhos da Comissão eram mulheres. Nalgumas comunidades, as mulheres não participaram nas actividades de socialização promovidas pela Comissão, visto ser suposto ficarem em casa. Além disso, havia proporcionalmente menos mulheres que eram membros de organizações formalmente constituídas com acesso às informações relativas ao trabalho da

* As equipas da Comissão coligiram um total de 7.824 testemunhos. Alguns (155 testemunhos) não foram introduzidos na HRVD porque não incidiam sobre violações no âmbito do mandato da Comissão ou porque não recaíam dentro do período de referência da Comissão.

† As Equipas distritais da Comissão trabalharam com as comunidades de acordo com a identificação local dos subdistritos, sucos e aldeias. Ao iniciar o seu trabalho no início de 2002, o número comum de subdistritos em Timor-Leste era de 65; Contudo, o Departamento Nacional de Estatística e o Levantamento de Sucos de Timor-Leste de 2001 registaram 64 subdistritos.

‡ A coligação de ONG incluía: o Centro de Serviços para Pessoas Internamente Deslocadas (*Center for Internally Displaced Persons Service*, *CIS*), *Truk-F*, a Organização de Advocacia para Oposição contra Violência sobre Civis (*Lembaga Advokasi Kekerasan Masyarakat Sipil*, *Lakmas*), *Yabiku* e a Fundação 'Apreensão pela Indonésia' (*Yayasan Peduli Indonesia*, YPI). Funcionários destas ONG recolheram testemunhos de timorenses de leste a viverem em Belu, Kefamenanu, Soe e Kupang em Timor Ocidental, entre Fevereiro e Agosto de 2003.

Comissão, e algumas mulheres manifestaram insegurança ou timidez em apresentarem os seus depoimentos.[*]

14. A Comissão recebeu testemunhos de adultos de todos os grupos etários. Tanto no caso dos homens como das mulheres, o número mais elevado de depoentes situou-se no grupo etário dos 40-44 anos, como se indica na Figura <g500002>.

[Insert Figure <g500002.pdf> about here]

15. Apesar das diferenças substanciais nas taxas de participação de homens/mulheres no processo de recolha de testemunhos da Comissão, os depoentes do sexo feminino falaram sobre as violações contra si (comparativamente às violações contra outros) numa proporção aproximadamente igual à dos depoentes do sexo masculino. Como se indica na Figura <tDepSexVictSexM>, de todas as violações reportadas por mulheres, 30,6% (2.939/9.605) foram violações contra a sua pessoa, enquanto que no caso dos depoentes do sexo masculino, 35,3% (17.438/49.382) das violações reportadas foram violações contra eles próprios.

[Insert Figure <tDepSexVictSexM> about here]

16. As barreiras sociais, culturais e económicas que as mulheres enfrentam podem ter limitado a sua participação nos processos de socialização e recolha de testemunhos promovidos pela Comissão. Contudo, as conclusões estatísticas da Comissão são consistentes com a afirmação de que a maioria das vítimas de assassinatos, desaparecimentos, tortura e maus tratos foram homens jovens do sexo masculino. Em contrapartida, a maioria esmagadora das violações de natureza sexual documentadas pela Comissão foram sofridas por vítimas do sexo feminino (ver Parte 6: Perfil das Violações de Direitos Humanos).

17. Os membros das equipas de recolha de testemunhos entrevistaram os depoentes em tétum, indonésio ou noutras línguas ou dialectos de Timor-Leste (que são línguas orais embora geralmente não escritas) e depois redigiram o texto da entrevista em tétum ou indonésio. Os formulários para recolha de testemunhos estavam disponíveis em tétum e indonésio. Dos 7.668 <s00101> testemunhos recebidos pela Comissão e considerados como cabendo no âmbito do seu mandato, 81,7% foram-no em tétum, 17,0% em indonésio, 1,2% noutras línguas de Timor-Leste e 0,1% numa língua que não foi especificada <s00002>. Uma vez que os formulários da Comissão para recolha dos testemunhos estavam redigidos em tétum e indonésio, os testemunhos prestados noutras línguas foram anotados pelos membros das equipas de recolha de testemunhos nos formulários oficiais em indonésio ou tétum antes de ser iniciado o processo de codificação, registo dos dados e análise dos testemunhos narrativos.

Enviesamento potencial das amostras no processo de recolha dos testemunhos

18. Como se refere na Secção abaixo, a natureza voluntária do processo de recolha de testemunhos da Comissão resultou em certa medida num processo de “auto-selecção”. Esta “auto-selecção”, por sua vez, introduziu um conjunto de factores que afectaram aqueles que estavam em condições de apresentar um testemunho, tais como:

[*] CAVR, documento interno: Evaluation Report of CAVR Statement Taking Process. Arquivo da CAVR.

- as pessoas que viviam em áreas remotas e montanhosas, muito longe dos locais onde os dados estavam a ser recolhidos (tais como as vilas dos distritos) tinham menor probabilidade de estar na amostra do que aquelas que viviam mais perto das vilas ou da capital do distrito.
- era maior a probabilidade dos testemunhos serem prestados por pessoas que eram socialmente activas e/ou fisicamente ágeis do que por aquelas que eram doentes, idosas, deficientes ou que estavam traumatizadas.
- era maior a probabilidade das pessoas que tinham um papel activo nas comunidades locais ou que estavam estreitamente ligadas a funcionários da administração pública local nas aldeias, subdistritos ou distritos e dos anciões participarem nos processos de socialização e de recolha de testemunhos, uma vez que essas iniciativas locais de recolha de testemunhos eram frequentemente organizadas através das estruturas e funcionários da administração pública local.
- as pessoas que faleceram antes da Comissão ser constituída não tiveram oportunidade para contar a sua história perante esta; por isso, houve uma tendência para os acontecimentos ocorridos no passado mais distante serem reportados com menos frequência do que os acontecimentos mais recentes.
- era menor a probabilidade de pessoas com pouco ou nenhum acesso aos órgãos e meios de comunicação contactarem a Comissão, e
- era menor a probabilidade das pessoas provenientes de grupos que eram hostis ao trabalho da Comissão apresentarem os seus testemunhos.

19. Para lidar com os problemas relacionados com o enviesamento da amostra, a Comissão complementou o seu processo de recolha de testemunhos com um conjunto de narrativas recolhidas pela *Fokupers* e informações de fontes secundárias da Amnistia Internacional. Além disso, para levar em linha de conta os desvios no cálculo das deslocações e violações fatais, a Comissão desenvolveu um Estudo Retrospectivo de Mortalidade que recolheu informação estruturada duma amostra probabilística aleatória de agregados familiares em Timor-Leste (ver mais à frente a secção que inclui uma apresentação detalhada da concepção das técnicas de amostragem e instrumentos de estudo usados no Estudo Retrospectivo de Mortalidade).

Amnistia Internacional

20. A Amnistia Internacional foi divulgando informações sobre a situação dos direitos humanos em Timor-Leste durante o período de referência da Comissão, sobretudo através de dados recolhidos pelas redes clandestinas no território e dos seus contactos com a diáspora timorense na Austrália e em Portugal.

21. A Comissão recebeu 322 relatórios e documentos da Amnistia Internacional, compilados entre 1975 e 1999.[*]

22. Os relatórios de natureza qualitativa da Amnistia Internacional e os seus documentos *Urgent Actions* (Acções Urgentes) foram codificados e introduzidos na Base de Dados das

[*] A Comissão não conseguiu localizar os seguintes relatórios da Amnistia Internacional:
ASA 21/12/83 UA 212/83 21 September
ASA 21/16/85 Disappearances
ASA 21/44/85 Unfair Trials and Possible Torture in Timor-Leste
ASA 21/22/87 Statement on ET by AI to the UN Special Committee on Decolonisation
ASA 21/23/87 ET: Releases of Political Prisoners
ASA 21/14/91 AI statement to UN Special Committee on Decolonisation - Appendix I and II
ASA 21/24/91 Timor-Leste: After the massacre – Appendix 1
Por esse motivo, a análise estatística que a Comissão faz das violações em Timor-Leste reportadas pela Amnistia Internacional não inclui actos relevantes referidos nestes relatórios.

Violações de Direitos Humanos da Comissão usando os mesmos métodos e normas utilizados para os testemunhos recolhidos pela Comissão. A informação recolhida pela Amnistia Internacional descreve a situação geral dos direitos humanos em Timor-Leste, tal como foi observada à época pela comunidade internacional das organizações defensoras dos direitos humanos.

Fokupers

23. O Fórum para a Comunicação entre as Mulheres Timorenses (*Forum Komunikasi Untuk Perempuan Loro Sae*, *Fokupers*), uma ONG de direitos humanos local, criou uma base de dados sobre violações na sequência da violência relacionada com a Consulta Popular de 1999.[*] A base de dados da *Fokupers* foi construída a partir de entrevistas abertas com mulheres timorenses conduzidas por pessoal da ONG. O objectivo original das entrevistas estava relacionado com o trabalho de aconselhamento realizado pela *Fokupers*. Contudo, esse objectivo foi alargado de modo a incluir documentação para efeitos de investigação pelas autoridades judiciais competentes, tais como a Unidade de Crimes Graves das Nações Unidas. Os testemunhos narrativos foram recolhidos em tétum.

24. A *Fokupers* construiu a sua base de dados para facilitar a publicação dum relatório sobre a violência contra mulheres. A base de dados original estava centrada na representação de dados biográficos sobre as vítimas, na narração dos acontecimentos, na identificação das violações cometidas e na identificação dos perpetradores. Em Julho de 2004, a *Fokupers* entregou estes dados à Comissão na condição dos depoentes, vítimas, membros das respectivas famílias referidos na base de dados não serem identificados no Relatório Final da Comissão. Os funcionários da Comissão registaram os dados, com base nas definições padronizadas e no esquema de codificação da Comissão, de modo a que os mesmos pudessem ser analisados em paralelo com a Base de Dados das Violações de Direitos Humanos da CAVR.

Codificação das fontes qualitativas (testemunhos narrativos da CAVR, Amnistia Internacional e Fokupers)

25. A codificação de dados é o processo através do qual a informação narrativa não estruturada sobre violações, vítimas e perpetradores é transformada num conjunto quantificável de dados, sem que seja descartada informação importante ou deturpada a informação recolhida.

26. Em Outubro de 2003, a equipa que na Comissão é responsável pelo processamento de dados reviu o processo de codificação e registo de dados para identificar erros sistemáticos e inconsistências no mesmo. Até essa data tinham sido codificados e introduzidos na base de dados da Comissão 2.473 testemunhos. Foi seleccionada uma amostra aleatória de 15% dos testemunhos (i.e., 371 testemunhos) da base de dados e estratificada em relação ao distrito em que o testemunho fora recolhido.

27. Cada testemunho foi revisto por um codificador: o codificador codificou de novo o testemunho sem olhar para a sua codificação original. Depois, os resultados de ambos os processos foram comparados e os erros na codificação original identificados, anotados e corrigidos. Além disso, o codificador pôde também rever o registo na base de dados para esse testemunho e identificar e anotar quaisquer erros na introdução dos dados, corrigindo-os subsequentemente.

28. Nos 371 testemunhos revistos foram identificados 416 erros de codificação. 58% (241/416) desses erros eram erros de codificação das violações, 12% (49/416) eram erros

[*] A *Fokupers* foi fundada em 1997 para apoiar as vítimas de violência política através de programas de aconselhamento e de outras formas de assistência às mulheres vítimas de violações, incluindo ex-presas políticas, viúvas de guerra, e mulheres de presos políticos. O mandato da ONG também inclui a promoção dos direitos humanos das mulheres entre a população local, especialmente entre as mulheres de Timor-Leste.

associados à codificação da filiação das vítimas, 10% (42/416) dos erros tinham que ver com o nível de especificação do(s) local(ais) codificado(s) e 9% (36/416) estavam associados à codificação da filiação institucional do perpetrador. Dos 416 erros de codificação identificados, 70% (291/416) eram erros de não identificação (ie, em que o acto não era identificado como uma violação ou a pessoa ou local não era identificado pelo codificador). Outros 17% (71/416) dos erros de codificação resultaram do facto do codificador incluir o acto como uma violação quando aquilo que era descrito na narrativa não cumpria as definições e condições-limite do vocabulário controlado instituído pela Comissão. Finalmente, 13% (54/416) dos erros de codificação eram o resultado da classificação errada dum acto numa categoria incorrecta de violações.

29. Na sequência desta revisão do processo de codificação, a equipa de processamento de dados desencadeou três iniciativas destinadas a minimizar no trabalho futuro os erros identificados: (1) foram feitas algumas revisões ao vocabulário controlado da Comissão; (2) foi organizado um seminário em que os resultados da revisão foram apresentados à equipa de codificação e onde foi dada formação adicional nas áreas consideradas necessárias; e (3) foram postos em prática exercícios regulares de codificação em grupo, em que os codificadores codificavam os mesmos testemunhos e reviam a consistência das suas decisões de codificação utilizando não apenas revisões qualitativas como medidas quantitativas de fiabilidade interavaliadores (*Inter-Rater Reliability*, *IRR*).[*]

30. Os principais tipos de revisões introduzidos no vocabulário controlado da Comissão foram:

- a redução do número de categorias de violações e a criação duma lista mais gerível
- as condições-limite foram afinadas para categorias de violações conceptualmente semelhantes (tais como tortura e maus tratos)
- recentragem do vocabulário controlado apenas na medição das violações, e não na medição destas e no seu impacte físico e psicológico
- simplificação das definições das categorias de violações, assegurando uma maior consistência entre a sintaxe da definição e a especificidade da informação recolhida nos testemunhos (por exemplo, os termos técnicos jurídicos foram reescritos em linguagem comum ou eliminados, uma vez que não se adequavam à realidade histórica que estava a ser medida)
- revisão da lista de actores institucionais; simplificação da lista e, simultaneamente, uma estruturação hierárquica das instituições que reflectisse as suas inter-relações estruturais.

Resultados da recolha de dados na HRVD

31. As três fontes de dados combinadas da HRVD produziram uma base de dados com registos tais como aqueles que se apresentam na Figura {ZZ} . Esses registos representam vítimas individuais e vítimas em grupos que sofreram violações fatais e não fatais. A Figura {ZZ} apresenta uma desagregação do número de registos recolhidos em cada base de dados. Note-se que esses números representam totais de dados antes das operações de limpeza em que os registos inválidos e duplicados foram removidos das bases de dados.

Table 1 - Figura {ZZ}: Matriz de contabilização de registos para a Base de Dados das Violações de Direitos Humanos

	Nº Testemunhos	Nº Indivíduos	Violações Fatais	Violações Não Fatais
Testemunhos CAVR				

[*] A fiabilidade interavaliadores (*Inter-Rater Reliability,* IRR) é uma medida do grau de concordância entre dois codificadores. A *IRR* serve para avaliar a consistência da implementação dum sistema de codificação.

	Nº Testemunhos	Nº Indivíduos	Violações Fatais	Violações Não Fatais
Amnistia Internacional				
Fokupers				
Totals				

32. Os grupos são registos de vítimas não nomeadas que identificam duas ou mais vítimas. Algumas vítimas sofreram violações não fatais múltiplas, outras sofreram violações não fatais e uma violação fatal, e outras ainda sofreram apenas uma violação fatal. Por conseguinte, o total de violações não coincide com o total de vítimas.

Estudo Retrospectivo de Mortalidade (*Retrospective Mortality Survey*, RMS)

33. A Comissão realizou um Estudo Retrospectivo de Mortalidade (*Retrospective Mortality Survey*, RMS) para obter uma estimativa de base probabilística do número de deslocações e de mortes. Este estudo partiu duma amostra aleatória estratificada de agregados familiares e recorreu a um questionário estruturado para recolher informação acerca de mortes nas famílias e deslocações durante o período de referência da Comissão. O estudo permitiu calcular estimativas estatísticas da mortalidade natural, mortes relacionadas com a fome, mortes relacionadas com o conflito, e migração.

Amostragem estatística usada no RMS

34. A amostra do RMS baseou-se num procedimento de amostragem em duas fases. A primeira fase foi uma amostra de todas as 2.336 aldeias de Timor-Leste, e a segunda fase uma amostra de agregados familiares em aldeias seleccionadas.[*]

35. A população dos agregados familiares foi estratificada de acordo com as seguintes variáveis: urbana/rural, localização do distrito, e altitude.[†] Foram usados métodos de estratificação implícita de modo que a lista de aldeias foi ordenada segundo as seguintes variáveis: urbanicidade, distrito e altitude, e uma amostra aleatória sistemática seleccionou amostras em cada uma das variáveis de estratificação.[‡] Foi criada uma medida de dimensão cumulativa e foi calculado um intervalo de amostragem como sendo o número de agregados de aldeias (*clusters*) (144) dividido pelo valor total da medida de dimensão (180,015), e que era igual a 1250,1. Foi gerado um número aleatório entre 1 e 1250,1 (397.235) e a aldeia com um valor da medida cumulativa de dimensão acima desse número foi seleccionada na amostra. Adicionou-se repetidamente 1250,1 ao número gerado aleatoriamente inicial e mais aldeias foram sendo seleccionadas da lista de modo análogo.

36. A decisão de estabelecer um número fixo de 20 agregados familiares, em vez de usar algo proporcional à dimensão da aldeia ou outro método de atribuição, deveu-se essencialmente a considerações de natureza operacional. A selecção de um número fixo de agregados

[*] A aldeia é a unidade administrative de menor dimensão em Timor-Leste. Em geral, uma aldeia é um conjunto de agregados familiares que partilham uma pequena área. Habitualmente, o suco é constituído por três ou quarto aldeias e um grupo de sucos forma o subdistrito que é uma divisão do distrito. De acordo com o Recenseamento de Sucos de Timor-Leste de 2001 existem 13 distritos, 64 subdistrictos, 498 *sucos*, e 2.336 aldeias em Timor-Leste. As equipas distritais da Comissão trabalharam nas 65 áreas consideradas subdistritos pelas comunidades, uma vez que foi necessário algum tempo para reorganizar as fronteiras administrativas após o fim da ocupação.

[†] A estratificação é o processo de agrupamento dos membros duma população em grupos relativamente homogéneos antes de se fazer uma amostragem. Os estratos têm de ser mutuamente exclusivos de modo a que cada elemento da população seja atribuído apenas a um estrato. Os estratos também devem ser colectivamente exaustivos, de modo a que nenhum elemento da população possa ser excluído. Uma vez definidos os estratos, aplica-se uma amostragem aleatória aos mesmos. As amostragens aleatórias estratificadas melhoram frequentemente a representatividade da amostra por reduzirem o erro de amostragem.

[‡] A Comissão utilizou um método conhecido como probabilidade proporcional à dimensão (*Probability Proportional to Size*) (neste caso a "dimensão" refere-se ao número de agregados familiares e não à população, embora os dois estejam obviamente relacionados), que é uma abordagem habitual em estudos deste tipo.

familiares por aldeia é uma forma de manter o controlo sobre a dimensão total da amostra e de garantir uma distribuição aproximadamente uniforme do volume de trabalho pelos entrevistadores.

37. A Comissão encarou a viabilidade de incorporar inquiridos timorenses ainda deslocados em Timor Ocidental na população de referência.[*] No entanto, questões de segurança, operacionais e de qualidade dos dados associadas às condições existentes em Timor Ocidental tornaram difícil a implementação do estudo nesse território. Por isso, a população de referência que foi considerada para a amostragem pela Comissão consistiu em todos os agregados familiares dos 13 distritos de Timor-Leste.

38. Por razões tanto estatísticas como operacionais, foi considerado não ser adequado incluir aldeias com menos de 20 agregados familiares no universo a ser objecto de amostragem. Por isso, as aldeias pequenas foram combinadas com aldeias próximas (mas não necessariamente adjacentes) antes de se fazer a amostragem, de modo que o número estimado de aldeias num agregado de aldeias (*cluster*) (definido como uma aldeia ou grupo de aldeias) era de pelo menos 40, para reduzir a probabilidade dalgum agregado na amostra ter menos de 20 agregados familiares. Na prática, e devido à inexactidão do universo em questão, uma equipa no terreno, ao chegar a uma aldeia, podia verificar que ela tinha menos de 20 agregados familiares, fosse porque o número de agregados reportados no censo de 1990 era incorrecto, ou porque a situação se alterara entretanto. Por esse motivo, os 144 agregados de aldeias incluídos na amostra continham, na realidade, 165 aldeias. Do ponto de vista operacional, isso significou que, nesses agregados de aldeias, os entrevistadores tiverem de criar uma amostra aleatória de 20 agregados familiares de entre o total combinado de agregados familiares no agregado de aldeias.

Concepção do questionário e desenvolvimento do Estudo Retrospectivo de Mortalidade

39. O questionário do RMS foi concebido para alcançar os seguintes objectivos:

- produzir estimativas da mortalidade total em Timor-Leste entre 1974 e 1999, usando não apenas técnica de estimação baseadas no estudo como técnicas de estimação de sistemas múltiplos (*Multiple Systems Estimation*, MSE), e
- desenvolver uma análise baseada no estudo que estimasse e descrevesse os complicados movimentos de deslocação no interior de Timor-Leste durante o período de referência da Comissão.

40. Assim, o questionário foi organizado nos seguintes módulos:

- um registo de agregados familiares
- um registo de deslocações de chefes de família
- uma história de nascimentos de mulheres adultas
- uma história de irmãos adultos homens/mulheres
- um história parental de adultos homens/mulheres
- uma secção genérica sobre violações de direitos humanos

41. O questionário[†] foi revisto por três estatísticos especializados em direitos humanos exteriores à Comissão[*] e por diversos especialistas em diferentes matérias da Comissão.

[*] O artigo 3º, nº 3 do Regulamento nº 2001/10 estabelece: "A Comissão pode realizar todas as actividades que forem consistentes com a prossecução do seu mandato tal como está definido no presente Regulamento."
[†] No Apêndice a este Anexo está reproduzido um exemplar do questionário do estudo.

Através desse processo de revisão, foram introduzidas melhorias no formato e concepção do questionário, tendo sido identificadas diversas questões terminológicas nas línguas indonésia e tétum.

42. Durante a fase de desenvolvimento do questionário foi realizada uma série de oito entrevistas cognitivas. As entrevistas cognitivas exploram os processos cognitivos do inquirido. Essas entrevistas procuram identificar as dificuldades e possíveis soluções para os desafios que os inquiridos enfrentam: (i) na compreensão das questões, (ii) na busca e recuperação da memória sobre a informação relevante, (iii) nos processos de tomada de decisões, e (iv) nos processos de elaboração das respostas.[†] Participaram nas entrevistas cognitivas um conjunto de oito indivíduos—quatro em condições de laboratório e quatro no terreno. Da análise da capacidade dos inquiridos para recordarem e recuperarem dados foram extraídas informações valiosas. Em particular, os processos cognitivos e as respostas a questões sobre tempo e datas indicaram que, muitas vezes, quando um inquirido respondia "Não Sei" podia apenas não saber a data exacta de acordo com o calendário Gregoriano. Contudo, as suas respostas indicavam que por vezes a localização no tempo dos acontecimentos era mais fácil de recordar por referência a outros marcadores de tempo tais como outros acontecimentos importantes, ou momentos particulares no ciclo agrícola ou das estações do ano.

43. A partir do processo de entrevistas cognitivas, foram desenvolvidos apontadores de datas (*date probes*) estruturados, que eram usados para solicitar aos inquiridos que restringissem as datas dos acontecimentos a uma janela de seis meses, a qual podia ser definida por acontecimentos importantes tais como feriados ou indicadores ambientais/físicos (altura do milho ou de outras culturas, estação das chuvas ou estação seca). O processo de entrevistas cognitivas também revelou que conceitos temporais tais como "início", "meio" e "fim" não eram entendidos por todos os inquiridos, pelo que não foi possível restringir mais as janelas temporais.

44. Durante as entrevistas cognitivas e no terreno, os inquiridos responderam com frequência simplesmente "Não Sei" ou "para as montanhas/floresta" como o local para onde tinham sido deslocados. Na sequência das entrevistas cognitivas, foi criado um conjunto de apontadores (*probes*) cuidadosamente elaborado para produzir descrições mais detalhadas dos locais para onde as pessoas tinham sido deslocadas.

45. Após revisão pelos pares e conclusão do processo de entrevistas cognitivas, o questionário finalizado foi traduzido e retrovertido para indonésio e tétum. O questionário foi depois testado no terreno durante 5 dias em aldeias de Díli que não faziam parte da amostra. Na sequência desse teste no terreno, foram introduzidas mais algumas alterações relativas à sequência das perguntas e melhoradas a gramática e sintaxe.

Implementação do estudo e trabalho de campo

46. Dentro de cada agregado familiar da amostra, o chefe de família respondeu tanto ao registo do agregado familiar (no qual eram registados todos os elementos do agregado familiar) como à secção sobre deslocações. Uma mulher adulta foi escolhida aleatoriamente de entre a população feminina adulta de cada agregado familiar para responder ao módulo sobre história de nascimentos de mulheres adultas.

47. Antes das equipas de entrevistadores abandonarem as aldeias, todos os questionários foram verificados por supervisores no terreno para identificar e corrigir quaisquer erros ou

[*] Fritz Scheuren, presidente da *American Statistical Association* , e consultor do HRDAG em projectos para o Kosovo, Guatemala e Peru; William Seltzer, Fordham University, e Jana Asher, co-autora dos relatórios do HRDAG sobre o Kosovo, Serra Leoa e Peru.

[†] Tourangeau 1984.

inconsistências no seu preenchimento. Dois coordenadores no terreno acompanharam a equipa de 22 enumeradores para o terreno.

48. Doze aldeias que tinham sido incluídas na amostra não puderam ser visitadas pela equipa de enumeração. A equipa não pôde realizar entrevistas nessas doze aldeias por causa da situação de segurança existente à época. A Figura {YY} indica as 12 aldeias que não foram enumeradas.

Distrito	Subdistrito	Suco	Aldeia
Alieu	Remexio	Liurai	Coto Mori
Baucau	Fatumaca	Samalari	Osso Luga
Baucau	Laga	Samalari	Soru Gua
Bobonaro	Atabae	Atabae	Heleso
Bobonaro	Bobonaro	Tapo	Tapo
Covalima	Fohorem	Datorua	Fatulidun
Lautém	Iliomar	Ailebere	Heitali
Lautém	Lospalos	Fuiluro	Kuluhun
Liquiça	Bazartete	Fahilebo	Fatu Neso
Oecusse	Passabe	Abani	Na Nos
Viqueque	Ossu	Uaibobo	Sogau
Viqueque	Uatu-Lari	Matahoi	Loko Loko

49. Além disso, em algumas ideias, foram enumerados menos de 10 agregados familiares, de que resultaram algumas não respostas adicionais. Globalmente, nos 1440 agregados familiares da amostra, houve uma taxa de não respostas de 3,1% (44/1440). Tendo em conta o valor reduzido da taxa de não respostas, não foi realizada nenhuma imputação estatística explícita para controlar as não respostas no estudo.

Base de Dados do Recenseamento de Cemitérios (*Graveyard Census Database*, GCD)

50. Para obter dados sobre a mortalidade de referência em Timor-Leste, a Comissão realizou um recenseamento dos cemitérios públicos nos 13 distritos de Timor-Leste. Através deste processo foi recolhida a informação disponível sobre nomes, datas de nascimento, datas de óbito e religião. As lápides que não dispunham dessa informação também foram enumeradas e a sua dimensão anotada.[*] Através da recolha dessa informação, a Comissão criou *de facto* um sistema de registo fundamental para a população timorense. Ou seja, a GCD criou uma listagem de referência de parte, talvez mesmo da maioria dos óbitos, que poderá ser usada para análises de mortalidade bem para lá deste projecto específico.

Recolha de dados para a GCD

51. Para facilitar o recenseamento que a Comissão pretendia realizar dos cemitérios públicos no país, foi elaborada uma lista de todos os cemitérios públicos conhecidos em Timor-Leste pelos funcionários da CAVR no terreno, em consulta com os funcionários da administração pública nas aldeias ao nível dos sucos e, onde tal foi possível, a nível das próprias aldeias. Um "cemitério público" neste estudo foi definido como um local reservado exclusivamente para enterrar pessoas falecidas. Esta definição inclui locais de enterro comunitários situados em terrenos públicos ou em terrenos propriedade de instituições religiosas. Contudo, a definição exclui sepulturas familiares localizadas em terrenos privados.

52. Os dados da GCD foram recolhidos por duas equipas de recolha de dados distintas. A primeira equipa recolheu 128.751 registos de 803 cemitérios, que foram introduzidos numa série de folhas de cálculo Excel. A primeira equipa cobriu partes de todos os 13 distritos, mas apenas

[*] A dimensão de uma lápide sem qualquer indicação pode ser usada como um indicador da presença duma criança ou adulto na sepultura.

o de Díli foi completamente abrangido pelo seu trabalho. Uma segunda equipa deslocou-se a todos os distritos, excepto ao de Díli, para concluir o recenseamento dos cemitérios. Essa equipa recolheu 153.057 registos adicionais de 1.779 cemitérios. A segunda equipa utilizou uma base de dados FoxPro para registar os seus dados.

53. As equipas de enumeração da Comissão documentaram todas as lápides existentes nos cemitérios públicos—tanto aquelas que continham inscrições como as que não continham. Uma sepultura contendo inscrições foi definida como possuindo uma estrutura física que recordava a vida duma pessoa falecida, com inscrições legíveis em inglês, indonésio, tétum ou português.[*] Em todas as lápides com inscrições enumeradas, a seguinte informação foi codificada, quando constava da lápide: nome completo, data de nascimento e data de óbito. As lápides sem inscrições eram tipicamente pequenas cruzes muito simples ou outros marcadores de sepulturas, sem nomes ou datas relativas à pessoa falecida. Aos enumeradores foi pedido que anotassem informações sobre a religião, tipo de material e dimensão das sepulturas, se tal fosse passível de identificação, tanto para as lápides com inscrições como para as lápides sem inscrições.

3. Descrição metodológica das técnicas de edição, limpeza e normalização de dados

54. Cada uma das três bases de dados usadas pela Comissão exigiu a aplicação de técnicas de edição e limpeza de dados e de normalização de nomes para que os dados pudessem ser comparados e ligados aos de outras bases de dados. Vários meses foram despendidos a rever os dados para detectar erros tipográficos ou ortográficos óbvios, e foi feita uma revisão duma amostra aleatória para garantir a sua exactidão. Houve problemas técnicos na conversão de dados de uma base de dados para outra, que também foram identificados e corrigidos.

Limpeza e edição das bases de dados

55. A equipa responsável pelo processamento dos dados fez uma verificação completa (e introduziu correcções, sempre que necessário) de todos os registos da HRVD em que:

- estava ausente de informação sobre o distrito/subdistrito
- havia informação não plausível sobre a data de ocorrência da violação (por ex, dia = 42, mês =13)
- a violação ocorrera antes da data de nascimento da vítima
- a violação ocorrera após a data de falecimento da vítima
- o inquirido fora codificado como vítima de uma violação fatal
- a idade da vítima fora registada como sendo 0 ou número negativo
- a idade da vítima fora registada como sendo superior a 75
- não fora inscrito qualquer código de violação
- não fora registada nenhuma vítima para uma violação codificada
- não existia qualquer perpetrador (individual/institucional) atribuído a uma violação codificada.

56. Para além das verificações rápidas acima descritas, a equipa de codificação também fez verificações numa amostra aleatória simples de registos de violações fatais, detenções, tortura, maus tratos, recrutamento forçado, violações de natureza sexual e deslocações. O objectivo das

[*] Por escassez de recursos, a Comissão não pôde enumerar as lápides chinesas.

verificações rápidas era identificar se existiam quaisquer erros sistemáticos na atribuição de filiações das vítimas e de responsabilidade institucional aos perpetradores. Foi identificada uma importante inconsistência – nomeadamente nos casos em que a filiação não era atribuída a todas as vítimas de uma violação ou violações ocorridas no mesmo acto ou actos muito próximos no tempo. Esses registos foram identificados e aplicadas as regras adequadas para atribuir correctamente a filiação das vítimas de todas as violações num mesmo acto ou de actos próximos para um mesmo actor.

Edição e limpeza das datas

57. Os registos que apresentavam erros óbvios, tais como datas de nascimento, violação ou falecimento posteriores à data de criação do registo foram examinados e corrigidos. Este problema era particularmente comum na base de dados GCD, onde os marcadores das sepulturas eram tão pequenos que não havia possibilidade de inscrever datas de anos completas com quatro dígitos. O sistema de registos de dados atribuía por defeito os dois dígitos em falta às datas, mas usando 20 em vez de 19 e colocando-as no século XXI. Os enumeradores de diferentes equipas usaram regras diferentes para a codificação das datas. Alguns recorreram à norma europeia DD-MM-AAAA, outros à norma americana MM-DD-AAAA", outros a um formato AAAA-MM-DD, ou variações destes usando dois dígitos para a representação do ano. Além disso, foram usados por vezes diferentes separadores entre anos, meses e dias – incluindo "/", ".", e "-". Por conseguinte, houve necessidade de converter os três conjuntos de dados no formato padronizado AAAAMMDD.

58. Se a data de nascimento (*Date of Birth*, DOB) estava depois da data de falecimento (*Date of Death*, DOD), as duas data eram trocadas. Também foram identificados e examinados dois tipos de erros que originaram datas com meses de valor superior a 12 ou com mais de 31 dias. A Comissão concluiu que alguns erros eram provocados por variações no formato das datas nas folhas de cálculo dos computadores usados para introduzir os dados.

59. Outros erros eram obviamente de natureza tipográfica. Os registos da HRVD e da RMS foram corrigidos revendo o material em suporte papel e aplicando correcções à base de dados. Para a base de dados GCD não houve tempo suficiente para rever os dados na fonte e, por isso, nos casos em que os erros não eram fáceis de corrigir, os valores nessa parte do campo de dados (mês ou dia) foram deixados em branco.

Edição e limpeza dos dados referentes à idade

60. Os dados referentes às idades foram examinados para detecção de possíveis erros tipográficos, por exemplo pessoas com mais de 100 anos de idade. As fontes desses registos foram revistas para verificação dos dados e as correcções necessárias foram introduzidas. Nos casos em que a data de nascimento (*Date of Birth*, DOB) e a data de falecimento (*Date of Death*, DOD) eram conhecidas, calculou-se a respectiva idade. O valor da idade na GCD foi calculado e foi gerado um novo campo para facilitar as correspondências.

Edição e limpeza de códigos de violação e relação

61. Os códigos das violações e relações foram revistos no contexto dos códigos da HRVD e da RMS para identificar aqueles que não eram válidos ou estavam em contradição com outros dados num registo individual (por exemplo, uma mulher que era codificada como pai). Os ficheiros fonte em suporte papel para estes registos foram revistos e as necessárias correcções introduzidas na base de dados.

Edição e limpeza dos dados referentes à localização geográfica

62. Os dados referentes às localizações geográficas recolhidos para as bases de dados RMS e HRVD foram codificados de acordo com as normas de geocodificação estabelecidas pelo Governo de Timor-Leste e aprovadas para serem usadas pela Comissão. As localizações foram divididas em quatro níveis administrativos—Distrito, Subdistrito, Suco, e Aldeia. Para os locais situados fora de Timor-Leste foram criados códigos para Timor Ocidental e Java e quando a localização era desconhecida foi usado um código distinto para esse efeito. A cada cemitério foi atribuído um código único, designado "id", para permitir a diferenciação entre cemitérios situados na mesma área geográfica.

63. A informação da GCD não foi recolhida com os códigos geográficos padronizados de Timor-Leste, pelo que houve necessidade de a converter para esses códigos-padrão.

Eliminação da duplicação de cemitérios e sepulturas na GCD

64. Diversos factores originaram a duplicação de registos relativos a sepulturas e cemitérios na base de dados GCD.

- Diferentes equipas de recolha de dados cobriram inadvertidamente o mesmo cemitério. Muitos cemitérios não possuíam indicação de nome, o que dificultou a identificação de registos duplicados unicamente pelo nome do cemitério.
- A localização exacta do suco e aldeia era frequentemente difícil de determinar em algumas zonas rurais. Mesmo que o cemitério tivesse o mesmo nome, podia ser codificado com uma localização geográfica diferente. Além disso, muitos cemitérios partilhavam o mesmo nome (sendo Santa Cruz o nome mais comum), o que significou que o nome do cemitério por si só não foi suficiente para identificar duplicações de cemitérios codificados com localizações geográficas diferentes.
- Muitos cemitérios em Timor-Leste não estão organizados de forma linear. Isto levou a que por vezes as equipas de enumeradores passassem duas vezes pela mesma sepultura, registando-a mais de uma vez.
- Por causa da enorme quantidade de registos em suporte papel necessários para recolher todos esses dados, é possível que tenha havido duplicações nas entradas dos registos.

65. Foi possível encontrar ligações entre ids de cemitérios através dum exame dos nomes das pessoas falecidas, localizações dos cemitérios, nomes dos cemitérios e datas completas de nascimento e óbito depois de estabelecida a correspondência.[*] Quando foram encontradas colunas de registos duplicados, um dos cemitérios foi eliminado do conjunto de dados para análise. Embora seja comum as pessoas terem o mesmo nome próprio e apelido, e possivelmente a mesma data de óbito, é altamente improvável que coincidam nas datas de nascimento e de óbito. Por isso, quaisquer registos que apresentassem o mesmo nome próprio, apelido, data de nascimento e data de óbito foram considerados duplicados, e apenas um deles foi mantido na base de dados para análise.

66. O objectivo do processo de eliminação de duplicações na GCD foi garantir que as pessoas falecidas eram contabilizadas apenas uma vez. Inicialmente pensou-se que durante as deslocações forçadas, as pessoas podiam ter sido inicialmente sepultadas nos locais onde haviam falecido, sendo os seus corpos posteriormente exumados pela família e sepultados na sua aldeia natal. Pensou-se também que se o corpo não fosse recuperado, seria erigido um

[*] Um registo completo é definido como um registo que possui informação sobre o dia, mês e ano, tanto para a Data de Nascimento como para a Data de Falecimento.

memorial no cemitério local. Embora isto possa ter acontecido, uma revisão cuidadosa dos dados não revelou serem o enterro num novo local ou a marcação *post hoc* práticas comuns. Além disso, nos casos em que os corpos foram recuperados, o mais provável foi a primeira lápide ser removida e transferida com o corpo, evitando desse modo uma duplicação das contagens. As pessoas que nunca foram sepultadas ou que não foram sepultadas em cemitérios públicos ficaram de fora da GCD. Para contabilizar as mortes que faltam nos depoimentos da HRVD, nas entrevistas da RMS, e nos dados sobre sepulturas da GCD, a Comissão realizou uma estimação de sistemas múltiplos (*multiple-system estimation*) do número total de mortes. Essa análise é descrita adiante.

Processos de limpeza de nomes

67. Os nomes das pessoas nos dados da Comissão tiveram de ser abordados de dois modos. Em primeiro lugar, os nomes tiveram de ser submetidos a uma análise gramatical (*parsing*) e distribuídos por três categorias—primeiro nome, nome do meio/alcunha e último nome. Uma vez concluído este processo, houve necessidade de converter os nomes numa forma canónica para facilitar a ligação dos registos (*record linkage*). Trata-se de um processo de redução de cada nome à sua forma mais simples e mais significativa, sem perda de generalidade.

68. Os nomes das pessoas apresentavam uma variabilidade considerável na sua ortografia, no modo como eram distribuídos pelos três campos definidos, e na pontuação. Existem muitas causas para a variabilidade dos nomes. Nos testemunhos narrativos abertos, tal como os que constam da HRVD, o depoente pode ser um familiar próximo, amigo, vizinho ou conhecido distante da vítima, e ele ou ela pode ou não saber como escrever os nomes da vítima reportada. A transcrição pela pessoa que recolhe o testemunho pode envolver a aplicação de regras ortográficas e de pontuação adicionais e mesmo introduzir erros ortográficos. Transformações análogas da ortografia e pontuação podem igualmente ocorrer quando os dados são codificados e registados nas bases de dados.

*Análise gramatical e distribuição os nomes por campos (*name parsing*)*

69. Para lidar com a variabilidade significativa no modo como os nomes eram atribuídos aos três campos – primeiro nome, último nome, nome do meio/alcunha, os nomes foram analisados gramaticalmente e distribuídos por campos (*parsed*) segundo regras estritas. O HRDAG decidiu dividir os nomes usando o "primeiro" primeiro nome para *primeiro nome*, e o "último" último nome para *último*, sendo todos os outros nomes colocados no campo *nome do meio/alcunha*. Além disso, as proposições (por exemplo, de, da, do, dos) foram eliminadas dos campos onomásticos, uma vez que sua utilização era inconsistente nos dados.

70. Por exemplo, o nome português Maria Luisa da Costa da Silva podia ter sido registado na base de dados como:

Primeiro nome	Nome do meio/Alcunha	Último nome
MARIA LUISA		DA COSTA DA SILVA
MARIA	LUISA	DA COSTA DA SILVA
MARIA LUISA	DA COSTA	DA SILVA
MARIA	LUISA DA COSTA	DA SILVA
MARIA LUISA		SILVA

71. O processo de análise gramatical e distribuição do nome por campos teria padronizado estas designações de modo a que o primeiro nome fosse Maria e o último simplesmente Silva. Todos os outros nomes, excluindo as preposições, seriam deslocados para o campo do nome do meio/alcunha.

72. O nome timorense indígena Mau Bere pode ter sido registado como:

Primeiro nome	Nome do meio/Alcunha	Último nome
MAU BERE		
MAUBERE		
MAU		BERE
		MAUBERE

73. O processo de análise gramatical do nome e sua distribuição por campos colocaria neste caso Mau no primeiro campo onomástico e Bere no último.

Transformação dos nomes na sua forma canónica

74. A transformação dos nomes na sua forma canónica foi aplicada aos campos dos registos contendo o primeiro e último nomes após a operação de análise gramatical acima descrita, a fim de facilitar o estabelecimento de correspondências, especialmente a utilização dos algoritmos automatizados para ligação de registos (*record linkage*). As variantes ortográficas dos nomes foram convertidas numa única forma representativa para cada nome. Por exemplo, as seguintes variantes ortográficas foram convertidas na forma canónica AGUSTINO:

- AGUSTINUHO
- AAGUSTINO
- AGUSTIO
- AGUSTINUS
- AUGUSTINHO
- AGUSTINO
- AGUSTINU
- AGUSTONIO
- AGUSRINO
- AGUSTINHO
- AGUSTIMHO
- AGSSTINHO
- AGSTINHO
- AUGUSTINO
- AGOSTINHO
- AGUASTINHO
- ANTGOSTINHO
- AGUSTINHU
- AGOTINHO
- AGOSTINO

75. Em relação aos nomes timorenses indígenas foi mais difícil encontrar formas canónicas, uma vez que têm em geral quatro ou cinco caracteres e alguns registos que parecem ser variantes ortográficas correspondem, na verdade, a nomes distintos. Em relação aos nomes timorenses indígenas aplicou-se um procedimento conservador para determinar as formas canónicas. O resultado foi depois testado com a ligação duma amostra de registos animistas,

tendo sido verificada a informação relativa à data, idade e local com vista a encontrar forma canónicas adicionais para aplicar.

76. Após diversas iterações do processo de análise gramatical dos nomes e sua distribuição por campos para encontrar as respectivas formas canónicas, foi gerado um novo campo com o nome grafado na ordem inversa. Em seguida, a ordenação alfabética deste campo permitiu encontrar nomes adicionais para serem convertidos numa única forma canónica, uma vez que as letras iniciais podiam variar, dependendo da pronúncia, mas a sílaba final tendia a manter-se idêntica. Este processo revelou-se muito útil na busca de formas canónicas adicionais.

77. Existiam nomes chineses, indonésios (muçulmanos) e anglo-saxónicos nas bases de dados, bem como nomes portugueses e nomes timorenses indígenas. O número relativamente pequeno de nomes chineses, indonésios e anglo-saxónicos não exigiu qualquer tratamento especial. O pessoal timorense da Comissão em Timor-Leste identificou quais os nomes que eram indígenas para efeitos da aplicação das regras e algoritmos de correspondência, uma vez que os nomes timorenses indígenas nem sempre são específicos quanto sexo daquele que nomeiam.

78. As bases de dados HRVD e RMS são mais pequenas do que a base de dados GCD, pelo que foram sujeitas primeiro ao processo de determinação de formas canónicas. Depois, a lista de nomes na sua forma canónica foi aplicada à GCD. Os nomes resultantes foram depois revistos para se identificarem formas canónicas adicionais.

79. Durante o processo de determinação de formas canónicas, descobriu-se que algumas letras nos nomes eram intermutáveis, sobretudo nos nomes portugueses. As letras S, J, G, e Z apareciam nos nomes muitas vezes nos lugares umas das outras. Também as letras V, U, W, e B eram frequentemente intermutáveis. Com menos frequência, havia trocas entre as letras H e E, ou então simplesmente a eliminação duma delas, como por exemplo Helder/Elder, Henrique/Enrique. Um exemplo dum nome com letras intermutáveis é, por exemplo, o do nome Virginia, que pode ser grafado com B ou V. Por exemplo, as variações ortográficas encontradas para a forma VIRGINIA incluem BIRGINIA, BERGINA.

80. Os nomes que começavam com estas letras foram comparados uns com os outros no âmbito do processo de determinação das respectivas formas canónicas. Nos casos em que os nomes possuíam mais de uma letra intermutável ou em que a letra intermutável estava no meio ou fim do nome, foi muito difícil encontrar potenciais formas canónicas. Por isso, foi escrito um programa que gerou uma lista de nomes onde as combinações de letras intermutáveis correspondiam à forma canónica doutro nome. O especialista em ligação de registos reviu essas combinações para determinar se deveriam ser transformadas em formas canónicas ou se correspondiam a nomes distintos e únicos. Nos casos em que existiam formas adicionais devido à presença de letras intermutáveis, a letra preferida para a forma canónica foi S (para S, J, G, e Z), V (para V, U, W, B), e H (para H e E).

81. Adicionalmente, no processo de determinação das formas canónicas, reparou-se que ANJU e ANJO eram frequentemente citados como primeiro nome ou como único nome dum registo. *Anju* é uma expressão comummente usada para fazer referência a uma criança falecida e foi encontrado com frequência nos registos da GCD, quando uma criança falecera antes de ser baptizada e portanto não recebera um nome cristão. Os registos com *ANJU* e um último nome foram usados para o processo de correspondências porque existiam alguns dados de identificação, mas os registos que apenas indicavam *ANJU* foram considerados demasiado ambíguos para permitir fazer sobre eles um juízo razoável com vista ao estabelecimento de correspondências.

Codificação do sexo e etnicidade

82. Durante o processo de determinação das formas canónicas, os primeiros nomes portugueses foram revistos em conjunto com as frequências dos códigos de sexo masculino, feminino e desconhecido.[*] Os códigos de sexo que estavam obviamente incorrectos foram corrigidos. Como sucede com a maioria dos nomes latinos, aqueles que terminam em A são geralmente femininos, e os que terminam em O (ou U) são geralmente masculinos. Quando os primeiros nomes terminavam em letras que não A, O ou U, as frequências respectivas de codificação masculina e feminina foram examinadas e quando existia uma grande disparidade, indicando que alguns registos tinham sido mal codificados durante o registo dos dados, fizeram-se correcções à base de dados.

4. Conversão de dados

83. Com vista a acelerar todos os passos do processamento dos dados associados ao estabelecimento de correspondências entre registos duplicados, cada conjunto de dados foi transferido da sua plataforma original FoxPro ou Excel para a plataforma de base de dados Analzyer.[†] O esquema da base de dados FoxPro foi primeiro duplicado em PostgreSQL para importação para Analzyer. As estruturas de bases de dados relacionais da HRVD e da RMS foram conservadas em Analzyer.

84. A Figura {XX} mostra o número total de registos de cada conjunto de dados que foram importados para Analzyer. De notar que estes totais reflectem as alterações resultantes das operações de limpeza dos dados que originaram a eliminação de registos duplicados e inválidos.

Table 2 - Figura {XX}: Nº total de registos por base de dados antes e depois da limpeza dos dados

Base de dados	Antes da limpeza	Após a limpeza	Nº/% Fatal	Nº/% Não Fatal
HRVD				
RMS				
GCD				n/a [‡]

5. Descrição geral das técnicas de ligação de registos (*record linkage*)

85. Os indivíduos referidos na HRVD e na RMS são por vezes referidos múltiplas vezes por diferentes depoentes, e podem igualmente surgir como registos na GCD. Para garantir que a análise estatística tinha em conta a existência de relatos duplicados sobre uma mesma pessoa, houve necessidade de fazer ligação de registos (*record linkage*) entre os dados, também conhecida como um processo de estabelecimento de correspondências (*matching*). Este procedimento foi aplicado a duas categorias genéricas de violações para efeitos de estudo – violações fatais e não fatais. As violações fatais incluem assassinatos de civis, mortes devidas a privações, desaparecimentos e mortes de combatentes. As violações não fatais incluem tentativas de assassinato de civis, prisões, torturas, violações sexuais, escravatura sexual, violência sexual, maus tratos, deslocações, casamentos forçados, impedimentos ao pleno exercício dos direitos reprodutivos, julgamentos não equitativos, destruição de casas, destruição de gado, extorsão, ameaças, recrutamentos forçados e trabalhos forçados.

[*] A frequência é uma contagem das instâncias em que um nome ou código surge num campo de dados particular. Os valores de frequências muito baixos podem revelar erros potenciais ou erros de ortografia nos dados.

[†] Analyzer é uma aplicação informática gratuita e de código aberto usada para recolher, manter e analisar informação acerca de violações de direitos humanos em grande escala. Para mais informações sobre a Analyzer, consultar o sítio internet do HDRAG: http://www.hrdag.org/resources/data_software.shtml.

[‡] Este campo não é aplicável no caso da GCD porque, por definição, uma pessoa sepultada num cemitério está morta.

86. Foram estabelecidos dois tipos de correspondências para efeitos de estimação estatística; correspondências intra- e inter-sistema (*intra-* e *inter-system matching*). As correspondências intra-sistema ligam registos que identificam a mesma pessoa dentro dum único conjunto de dados, e cada registo pode corresponder a zero, um ou muitos outros registos dentro desse conjunto de dados. As correspondências intra-sistemas ligam duas ou mais listas de registos únicos de diferentes fontes de dados, para que seja possível fazer uma estimativa das violações pelo método MSE. Os registos que são feitos corresponder durante um procedimento de estabelecimento de correspondências inter-sistema só podem corresponder a zero ou a um outro registo em cada um dos outros conjuntos de dados.

87. Por causa da complexidade inerente ao processo de estabelecimento de correspondências inter-sistemas e das limitações temporais ao trabalho a realizar, os dados relativos a violações não fatais constantes da HRVD e do RMS só foram submetidos a um processo de estabelecimento de correspondências intra-sistema para efeitos de estatística descritiva. Os dados relativos às violações fatais, que incluem os dados da GCD, foram submetidos a um estabelecimento de correspondências tanto intra- como inter-sistema como base para o cálculo de estimativas pelo método MSE. O estabelecimento de correspondências foi realizado usando três métodos: estabelecimento de correspondências "à mão", gerado por computador e assistido por computador. Cada um destes métodos pode envolver mais de uma iteração.[*]

Regras de correspondência

88. Cada registo individual foi comparado com todos os outros registos de cada conjunto de dados para identificação de correspondências possíveis, e foi considerado como correspondente quando um número significativo dos valores dos campos correspondiam *exactamente*, apresentavam uma *proximidade estreita*, ou *não estavam em conflito*. Os campos usados no processo de estabelecimento de correspondências foram: primeiro_nome, último_nome, idade, sexo, data de nascimento (*Date of Birth*, DOB), data de falecimento (*Date of Death*, DOD), local_de_nascimento (*Place of Birth*, POB), e local_de_falecimento (*Place of Death*, POD). Os campos do nome do meio/alcunha e local_da_entrevista também estavam disponíveis para efeitos de esclarecimento, mas não eram campos disponíveis nos três conjuntos de dados, e quando existiam, eram campos esparsamente preenchidos. Embora não fizessem parte das regras de correspondência, esses dados foram tidos em conta pelo especialista em ligação de registos. Contudo, não foram usados em qualquer procedimento de autocorrespondências computorizado.

89. As decisões relativas ao estabelecimento de correspondências usadas para os dados da Comissão tendiam a criar uma correspondência excessiva (*over-match*) dos registos.[†] Essa correspondência excessiva reduz o número de registos únicos e, por isso, tende a baixar o valor das estimativas. Este efeito de correspondências excessivas é preferido nos casos em que existe incerteza acerca da precisão duma correspondência, a fim de produzir estimativas conservadoras.

Correspondência dos nomes

90. Os campos do primeiro e último nome não estavam sempre preenchidos; alguns apresentavam iniciais ou faltava o primeiro ou último nome. Foram feitas tentativas para fazer corresponder cada registo, mesmo quando estava incompleto, mas nos casos de correspondências para violações fatais, os registos sem o primeiro ou último nomes ou apenas

[*] Uma iteração é uma revisão de todos os dados num conjunto de dados baseada na ordem pela qual foram ordenados ou no algoritmo de ordenação, a fim de procurar correspondências.

[†] A correspondência excessiva (*over-matching*) significa que podem ser estabelecidas ligações entre registos que podem não ser, na verdade, duplicados.

com iniciais foram retirados do processo de estabelecimento de correspondências uma vez que não existiam dados suficientes para fazer juízos fiáveis. Nos casos das correspondências para violações não fatais, foram feitas tentativas para fazer corresponder as violações com DOB, DOD, e informação sobre o local da morte a outros registos com idênticos valores nesses campos, mesmo quando não existia nome ou o registo apresentava apenas iniciais. Os registos com dados menos completos sobre os nomes dependiam mais fortemente de datas e locais perfeitos para poderem ser feitos corresponder a outros registos. Muitas pessoas podiam ter morrido no mesmo dia no mesmo local, e é difícil e pouco seguro saber a quais dessas pessoas se deve fazer corresponder um nome incompleto.

Correspondência do sexo e origem étnica

91. Nos casos em que o sexo da vítima era conhecido, só podia potencialmente ser feito corresponder a registos do mesmo sexo ou àqueles em que o sexo era desconhecido. Os registos em que o sexo era indicado como Desconhecido podiam ser feitos corresponder a registos com códigos Masculino ou Feminino, mas dentro dum grupo onde se tinham estabelecido correspondências os códigos relativos ao sexo não podiam estar em conflito com outros registos nesse grupo.

Correspondência das localizações

92. Os códigos de localização geográfica usados para os dados da CAVR foram divididos em quatro níveis: distrito, subdistrito, suco e aldeia. A base de dados GCD era o único conjunto de dados que apresentava a informação sobre a localização de forma desagregada até ao nível da aldeia, pelo que não foi usada para efeitos de estabelecimento de correspondências. A frequência das deslocações tornou difícil às testemunhas localizar com precisão os locais relevantes, excepto nos casos em que a violação ocorrera no local onde a testemunha residia na altura ou de onde fora originalmente deslocada. As pessoas podiam ter sido deslocadas múltiplas vezes, de e para diversos locais, e uma vez que o conflito se prolongou por três décadas a rememoração dos locais exactos estava sujeita a diversos erros.

93. Além disso, os limites entre localizações geográficas são afectados por três factores—alterações nos nomes dos locais e dos limites geográficos das divisões administrativas ao longo do tempo; imprecisão nos limites, especialmente nas zonas rurais; e erros potenciais na recolha dos dados, codificação e registo dos mesmos nas bases de dados. Por conseguinte, foram consideradas as correspondências entre quaisquer pontos dentro dum mesmo distrito e entre distritos adjacentes. Às correspondências potenciais entre um subdistrito e um suco que estavam mais perto um do outro também foi atribuída uma preferência mais elevada. Na análise pormenorizada dos dados, os registos que apresentavam predominantemente correspondências em campos de dados que não os referentes à localização, serviram para substanciar as nossas decisões relativamente às correspondências entre localizações. Nos casos em que a HRVD documentava uma morte ocorrida no mesmo local onde se realizara a entrevista, partimos do princípio que a informação sobre a localização tinha grande probabilidade de estar correcta.

94. Em casos raros, foram estabelecidas correspondências que violavam a regra relativa aos dados sobre localização, mas apenas quando era por demais evidente que os registos identificavam a mesma pessoa, e que erros tipográficos comuns eram responsáveis pela diferença verificada. Quando existia mais de uma possibilidade de correspondência, o algoritmo utilizado procurou estabelecer a correspondência com registos menos específicos, de modo a conservar os registos mais específicos para candidatos ulteriores. Quando existia uma distribuição igual entre localizações em qualquer nível geográfico, foi dada preferência à

localização menos específica, e quando não havia alguma que fosse mais ou menos específica, foi escolhida aleatoriamente uma dessas localizações para ser a "rep rec".[*]

Correspondência de datas

95. Uma vez que os conflitos em Timor-Leste se prolongaram por um extenso período, muitos dos inquiridos não se recordavam das datas e locais exactos em que determinados acontecimentos tinham ocorrido. Considerou-se que os dados da GCD eram mais exactos no que se referia à informação sobre datas e locais, uma vez que os corpos teriam sido normalmente sepultados pouco tempo após ocorrer a morte, e geralmente perto do local onde esta tinha ocorrido. Ao procurar estabelecer correspondências no campo da data, o especialista em ligação de dados estabelecia ligações entre registos que distavam uns dos outros de mais ou menos três anos. As excepções a esta regra foram raras, e apenas utilizadas quando os outros campos de dados apresentavam fortes correspondências. Os registos com dados relativos a meses e dias apresentavam com frequência imprecisões na HRVD e no RMS, uma vez que a memória humana tende a ser falível para períodos tão longos. Por conseguinte, as datas mais específicas foram postas em correspondência umas com as outras nos casos em que estavam próximas, e em correspondência com datas menos específicas nos casos em que não estavam próximas.

Restrições às correspondências ao nível dos registos

96. Foram impostas determinadas restrições às correspondências a fim de evitar fenómenos de correspondências excessivas (*over-matching*). Mais especificamente, as seguintes correspondências não eram autorizadas:

- Registos de vítimas dum mesmo testemunho (porque cada testemunho identificava vítimas únicas que poderiam ter nomes idênticos por causa das suas ligações familiares)
- Dois acontecimentos não fatais não podiam ser feitos corresponder um ao outro se tivessem sido reportados no mesmo registo-fonte (porque os métodos de codificação dos dados e de representação das bases de dados utilizados não permitiam que fossem introduzidos nas bases de dados registos duplicados provenientes dum único testemunho)
- Um depoente não podia ser feito corresponder a uma violação fatal
- Um registo não fatal não podia ser feito corresponder a um registo fatal se quaisquer datas associadas às violações não-fatais se situassem antes da DOB do registo fatal
- Um registo não fatal não podia ser feito corresponder a um registo fatal se quaisquer datas associadas às violações não fatais se situassem após a DOD do registo fatal.

Correspondências intra-sistema

97. Dentro dum conjunto de dados, uma pessoa pode ser identificada por testemunhas múltiplas. O estabelecimento de correspondências intra-sistema permite ligar registos que identificam a mesma pessoa, a fim de gerar uma lista de pessoas nomeadas únicas e evitar uma contagem excessiva e, desse modo, estimativas por excesso. O estabelecimento de

[*] O "rep rec" é o registo que representa melhor o agrupamento de registos entre os quais foi estabelecida correspondência por ser aquele que apresenta os dados mais completos. Os registos com a data ou localização mais comum dentro desse grupo ou um registo com uma localização ou data mais precisos são considerados mais completos. Quanto mais completos forem os dados, melhor será cada iteração subsequente para estabelecimento de correspondências intra- e inter-sistemas. Uma vez que se estava a ligar os registos uns aos outros e a conservar os dados específicos de cada registo, em vez de apagar os registos duplicados, houve necessidade de analisar a variação entre os registos que tinham sido feitos corresponder uns aos outros para verificar se as diferenças poderiam alterar significativamente a análise.

correspondências intra-sistema é muito complexo e difícil de realizar numa base de dados, uma vez que uma pessoa pode ser feita corresponder com *n* outros registos no conjunto de dados. Por isso, os dados são manipulados numa folha de cálculo que torna mais fácil ordenar e reordenar os dados de modos múltiplos, a fim de localizar as ligações que têm de ser feitas.

98. O estabelecimento de correspondências intra-sistema num conjunto de dados antes da fusão dos seus registos com outros conjuntos de dados pode revelar padrões inerentes a esse projecto de recolha de dados. Alguns desses padrões podem ser erros sistemáticos na recolha dos dados, sua codificação ou introdução na base de dados, ou podem resultar da estrutura do processo de recolha dos dados. A observação de padrões dentro de cada conjunto de dados permite a investigação e, se for caso disso, a correcção dos erros subjacentes.

99. Se fossem combinados, os três conjuntos de dados da Comissão seriam demasiado grandes para se poder fazer uma operação de estabelecimento de correspondências de alta qualidade, uma vez que alguns dos padrões não teriam sido detectáveis por um observador humano. Ou seja, se os três conjuntos de dados fossem combinados numa lista única, a lista resultante teria mais de 160 000 registos. Encontrar correspondências entre registos numa lista desta extensão teria sido muito difícil para um observador humano.

Estabelecimento de correspondências entre violações fatais intra-sistema na HRVD

100. Em primeiro lugar, foram estabelecidas correspondências intra-sistema entre os dados relativos a violações fatais na HRVD, a fim de ligar registos que descrevessem a mesma vítima. Os registos foram importados para uma folha de cálculo e ordenados pelo primeiro nome, último nome, POD, e DOD, para encontrar registos que se correspondessem.

101. À medida que os registos eram ligados, foi escolhido um “rep rec”. Após cada operação de ordenamento, foi realizada uma iteração do procedimento de estabelecimento de equivalências e os registos ligados num grupo de correspondências foram escondidos (mas não eliminados) do ficheiro de dados extraído, deixando apenas o seu “rep rec”. Isto permitiu reduzir o *ruído* nos dados. O ruído pode ser definido como os registos não “rep rec”, num grupo de correspondências que distraiem aquele que estabelece as correspondências das relações potenciais que podem existir entre o “rep rec” e outros candidatos ao estabelecimento de correspondências. Quanto mais pequena for a lista de registos únicos, mais fácil se torna identificar as correspondências potenciais e outros padrões nos dados. Cada iteração subsequente do procedimento identifica correspondências adicionais e finalmente é obtida uma lista de registos únicos da totalidade do conjunto de dados. É feito um mínimo de cinco iterações do procedimento de estabelecimento de correspondências para cada conjunto de dados.

102. Os 15.043 registos de violações fatais do conjunto de dados da HRVD foram reduzidos a uma lista de 11.145 vítimas únicas. Todos os registos foramn depois importados de novo para o sistema de estabelecimento de correspondências entre dados Analzyer. Os registos que foram postos em correspondência foram ligados de novo ao “rep rec” para análise uma vez concluídas todas as operações de estabelecimento de correspondências.

Estabelecimento de correspondências entre violações fatais intra-sistema no RMS

103. A operação de estabelecimento de correspondências entre violações fatais intra-sistema no RMS foi realizada numa folha de cálculo após conclusão da operação de estabelecimento de correspondências intra-sistema na HRVD. O procedimento de estabelecimento de correspondências intra-sistema no RMS recorreu aos mesmos campos que o procedimento de estabelecimento de correspondências intra-sistema na HRVD e também analisou a fonte do registo. Não foi permitida a correspondência de registos de fatalidades recolhidos do mesmo agregado familiar, uma vez que identificavam indivíduos únicos, mesmo que estes partilhassem o mesmo nome e DOD.

104. Os 4.883 registos de violações fatais do conjunto de dados do RMS foram reduzidos a uma lista de 4.619 vítimas únicas.

105. As ligações entre registos resultantes, tanto do conjunto de dados da HRVD como do RMS, foram importadas de novo para o modelo de dados Analzyer, a fim de serem usadas em procedimentos de estabelecimento de correspondências assistidos por computador e gerados por computador, e para gerarem dados para análise. As informações e os padrões documentados pelo especialista em ligação de dados na fase de estabelecimento de correspondências "à mão" foram depois usados para produzir regras e algoritmos de correspondência para os processos de estabelecimento de correspondências assistido por computador e gerado por computado.

Estabelecimento de correspondências entre violações não fatais intra-sistema na HRVD

106. Foram concebidos algoritmos informáticos para limpar e estabelecer correspondências entre violações não fatais HRVD. Este passo é referido como de *autocorrespondência* (*auto-matching).* Foram desenvolvidos algoritmos automáticos de estabelecimento de correspondências para as violações não fatais na HRVD, uma vez que as limitações de tempo e recursos não permitiram a utilização dum especialista humano em ligações de dados. Existiam três vezes mais vítimas de violações não fatais do que vítimas de violações fatais na HRVD.

107. A HRVD continha 41.546 registos. O procedimento de autocorrespondência intra-sistema produziu uma lista de 37.651 vítimas únicas de violações fatais e não fatais.

*Autodefinição de formas canónicas (*auto-canonicalisation*) de nomes de violações não fatais e estabelecimento de correspondências*

108. O primeiro passo na autodefinição de formas canónicas (*auto-canonicalisation*) foi a construção duma tabela com as diferentes versões limpas de todos os nomes originais (de violações fatais e não fatais) constantes da base de dados. Para o primeiro nome, as versões foram *normalizadas*, *normalizadas-modo* terse (*normalised-terse*), *normalizadas de primeiro nome* (*first word of normalised*, designadas *first-namefirst*), e *normalizadas de primeiro nome-modo* terse (*first word of normalised-terse*, designadas *first-namefirst-terse*). O mesmo método foi aplicado ao último nome, com a excepção de se ter usado a última palavra em vez da primeira palavra. Depois, para cada nome duma vítima de violação não fatal, foi feita uma tentativa de fazer corresponder as seguintes combinações dos nomes completos das vítimas de violações não fatais normalizados com todos os nomes completos de vítimas de violações fatais normalizados e cujas formas canónicas tinham sido definidas "à mão":

- namefirst + namelast
- namefirst-terse + namelast-terse
- first-namefirst + last-namelast
- first-namefirst-terse + last-namelast-terse

109. O programa de estabelecimento de correspondências operou sobre um conjunto completo de informação antes de tentar estabelecer correspondências num conjunto contendo um menor volume de informação. Estas correspondências entre nomes de violações não fatais e nomes de violações fatais foram apenas estabelecidas para nomes de violações fatais normalizados que correspondiam a um único nome canónico. À medida que a informação se foi tornando mais concisa (*terse*), existiam cada vez menos nomes normalizados "permitidos" para estabelecer correspondências (o que era contrabalançado pelo facto de ser mais fácil estabelecer as correspondências, uma vez que a informação menos concisa era mais resistente à variabilidade na codificação e aos erros na introdução de dados).

110. No caso dos nomes completos que não puderam ser transformados numa forma canónica, foram definidas formas canónicas independentemente para os primeiros nomes e os últimos nomes. A ordem de estabelecimento de correspondências para primeiros nomes foi a seguinte:

- namefirst
- namefirst-terse
- first-namefirst
- first-namefirst-terse

111. Foi depois desenvolvido um processo de estabelecimento de correspondências para se seguir ao procedimento preliminar de estabelecimento de correspondências baseado nos processos de autolimpeza (*auto-cleaning*) e autocorrespondência (*auto-matching*). Este processo destinava-se a correspondências potenciais com os nomes não normalizados e identificou a densidade de informação por campo de dados de cada registo de nome. A percentagem de registos contendo valores não vazios para os respectivos campos de dados era a seguinte:

- 9% tinham *date_birth* (todos estes tinham *birth_geo1*)
- 44% tinham birth_suco_location
- 50% tinham birth_subdistrict_location
- 53% tinham birth_district_location
- 70% tinham *Firstname*
- 94% tinham *Sex*
- 100% tinham *Lastname* (uma vez que se trata dum campo obrigatório para o estabelecimento de correspondências)

112. Uma vez que o campo do último nome era o único campo não vazio para todos os registos, era o único campo que podia ser usado para o procedimento de bloqueamento do índice (*index blocking*). Este procedimento de bloqueamento analisa os registos em que o(s) campo(s) que estão a ser bloqueados partilham o mesmo valor. O bloqueamento para o campo do último nome foi feito sobre as primeiras quatro letras de cada nome. O algoritmo de correspondência teve de ser cuidadosamente calibrado: se existissem muitos campos vazios, era necessária uma correspondência mais apertada nos campos não vazios (além disso, as correspondências no caso de últimos nomes muito comuns receberam uma ponderação menor).

113. Existiam três tipos diferentes de “proximidade” que foram feitos variar:

7. O número de letras no nome em que havia correspondência (4, 8, ou todas).
8. O número de níveis no local de nascimento em que havia correspondência (de 1 a 3), e
9. A proximidade exigida das datas (de 1/3 ano a 3 anos).

114. Com dois terços dos nomes das vítimas com formas canónicas autodefinidas, e um conjunto de regras bem definidas para a “proximidade de correspondência exigida” para diferentes números de campos não vazios, a taxa de correspondências resultante era de aproximadamente 15% (em comparação com cerca de 25% para os dados relativos às violações fatais cuja correspondência for a obtida por intervenção humana).

115. Uma taxa de correspondências de 15% para as violações não fatais parece plausível na medida em que:

- Apenas dois terços dos registos de nomes puderam ser reduzidos às suas formas canónicas, e
- É de esperar uma densidade de reporte mais elevada para as violações fatais, uma vez que estas são mais facilmente identificáveis e mais fáceis de rememorar por um maior número de pessoas na rede social da vítima.

116. O procedimento automatizado de estabelecimento de correspondências inter-sistemas aplicado às violações não fatais reduziu o conjunto de dados de 44.203 registos para uma lista de 1.568 registos únicos de vítimas

Revisão pelo especialista em ligação de dados das correspondências intra-sistema para as violações não fatais presentes na HRVD

117. O especialista em ligação de dados estudou uma amostra dos resultados das autocorrespondências para se assegurar de que não existiam falsas correspondências óbvias (ie. correspondências excessivas). Não foi detectado qualquer padrão sistemático de correspondências excessivas na revisão realizada a uma amostra aleatória de10% dos registos dos grupos de correspondências. O maior grupo de registos que tinham sido feitos corresponder uns aos outros incluía 20 registos. Foi feita uma revisão dos grupos maiores para garantir a plausibilidade da sua dimensão.

118. O estabelecimento de correspondências intra-sistema para os dados relativos às violações fatais gera uma lista combinada de indivíduos únicos que estão todos mortos, embora a causa da sua morte possa variar. Quando se utiliza o procedimento de estabelecimento de correspondências intra-sistema com as violações não fatais, uma vítima pode sofrer uma ou mais violações, em um ou mais dias, em um ou mais locais. No caso das violações não fatais, o estabelecimento de correspondências revela as violações de direitos humanos sofridas pelas vítimas individuais, sendo que uma vítima pode ter sofrido outras violações que podem ou não ter resultado na sua morte.

Correspondências inter-sistemas

119. O procedimento de estabelecimento de correspondências inter-sistemas associa listas de indivíduos únicos de conjuntos de dados múltiplos e é realizado cumulativamente em pares ou conjuntos de dados. O estabelecimento de correspondências inter-sistemas é aplicado apenas aos dados relativos às violações fatais. Primeiro, o estabelecimento de correspondências inter-sistemas é aplicado usando os 11.126 registos que foram feitos corresponder intra-sistema na HRVD e os 4.619 registos que foram feitos corresponder intra-sistema no RMS na aplicação Analzyer Record Linkage. O conjunto de dados *fonte* sobre violações fatais no RMS foi feito corresponder ao conjunto de dados *alvo* sobre violações fatais na HRVD.[*] [†]

Fase 1 – Estabelecimento de correspondências geradas por computador

120. O estabelecimento de correspondências estrito (designado estabelecimento de correspondências P1) identificou automaticamente "correspondências exactas". O processamento das "correspondências exactas" através do processo automatizado P1 elimina a

[*] Existe uma "correspondência exacta" nos casos em que dois ou mais registos numa base de dados são feitos corresponder e todos os campos sobre os quais são tomadas decisões de correspondência são idênticos.

[†] As designações de fonte e alvo são determinadas pelo número de registos no conjunto de dados. O mais pequeno dos dois conjuntos de dados do par é a fonte e o maior é o alvo. Isto destina-se a reduzir o número de registos que têm de ser comparados, embora cada registo de ambos os conjuntos de dados seja comparado com todos os seus correspondentes potenciais.

ineficiência associada a um processo em que um ser humano compararia cada registo nas, ou entre as, bases de dados, com todos os outros registos.

121. Foram aplicados algoritmos de estabelecimento de correspondências aos dados para produzir uma lista de correspondências potenciais consideradas altamente prováveis. Foram feitos cálculos baseados nas probabilidades e frequências de cada campo de dados dentro dum registo, os valores foram ponderados e ordenados, tendo sido estabelecido um nível-limiar a partir do qual se considerou que a correspondência feita estava provavelmente correcta. Esse valor-limiar foi estabelecido depois de uma revisão das correspondências prospectivas obtidas com o algoritmo, que eliminou a necessidade de intervenção humana para comparação de todos os registos em busca de possíveis correspondências. As correspondências potenciais abaixo do valor-limiar foram tratadas de dois modos, dependente do facto de se tratar duma correspondência para uma violação fatal ou não fatal, intra-sistema ou inter-sistemas.

122. No caso das correspondências inter-sistemas para dados sobre violações fatais, os conjuntos de correspondências gerados pelo algoritmo foram importados para o sistema de correspondência de dados Analyzer e o especialista em ligação de dados reviu estas correspondências assistidas por computador para cada um dos registos-fonte remanescentes que não tinham sido feitos corresponder. O procedimento de estabelecimento de correspondências intra-sistema para as violações não fatais foi completamente automatizado, sendo os resultados revistos pelo especialista em ligação de dados para garantir a não ocorrência de situações extremas de correspondências excessivas ou em número diminuto.

Fase 2 – Estabelecimento de correspondências assistido por computador

123. O estabelecimento de correspondências assistido por computador, designado P2, baseou-se em algoritmos que geravam bolsas de correspondências potenciais entre registos-fonte e registos-alvo que eram consideradas correspondências prováveis, mas para as quais era necessária uma intervenção humana de revisão para seleccionar qual dos registos cuidadosamente ponderados constituía a melhor correspondência. Foram feitos cálculos baseados nas probabilidades e frequências de cada campo de dados entre pares de registos, os valores foram ponderados e ordenados com base nos nomes, data de nascimento, data de falecimento, local de nascimento e local de falecimento. Usando a interface de correspondências do Analyzer, o especialista em ligação de dados seleccionou o registo-alvo da bolsa, quando existia, que correspondia ao registo-fonte que estava a ser examinado.

124. As regras para o estabelecimento de correspondências inter-sistemas para violações fatais P2 eram:

1. O sexo da fonte e alvo(s) tinham de ser iguais, quando o sexo era conhecido.
2. As primeiras iniciais dos nomes entre uma fonte e alvo(s) tinham de ser idênticas.
3. Para o(s) alvo(s), nos casos em que DOB e DOD eram conhecidas, uma das datas tinha de estar a uma distância não superior a 5 anos das datas da fonte.
4. Se a fonte e o(s) alvo(s) potencial(ais) tinham DOB ou DOD “perfeitas”, pelos menos um dos outros campos para correspondência tinha realmente de corresponder.

125. Depois de ter sido realizado o trabalho de estabelecimento de correspondências inter-sistemas em Analzyer entre os conjuntos de dados da HRVD e do RMS, a lista resultante de vítimas únicas de violações fatais foi importada para uma folha de cálculo. Os registos foram depois ordenados nos vários campos de dados para se determinar se era possível encontrar quaisquer outras correspondências. Esta operação serviu não apenas para apanhar correspondências que tinham escapado, mas também para avaliar o desempenho dos algoritmos de estabelecimento de correspondências. Na sequência das revisões feitas “à mão” pelo especialista em ligação de dados, os algoritmos foram afinados garantindo desse modo um

desempenho mais completo e preciso do algoritmo em iterações sucessivas do procedimento de estabelecimento de correspondências.

Fase 3 – Correspondência de dados vagos

126. Na Fase 3 de estabelecimento de correspondências (P3), foram estabelecidas correspondências entre registos que possuíam demasiados campos vazios, ou que eram registos de indivíduos com nomes comuns, duma mesma área geográfica, ou que tinham falecido no mesmo período temporal. No caso destas correspondências, não existiam dados suficientes para se especificar qual o par fonte/alvo que era correcto. Assim sendo, foi escolhido aleatoriamente um dos alvos. Por exemplo, Mau Bere era um nome muito comum e muitas partes do território, e 1999 foi um ano em que muitos deles morreram. É improvável que tenham existido correspondências intra-sistema que tenham escapado, por duas razões. Em primeiro lugar, tratava-se de registos que provinham frequentemente do mesmo testemunho, o que indicava tratar-se de membros duma mesma família com o mesmo nome. Em segundo lugar, a GCD regista muitas mortes num mesmo cemitério com o mesmo nome e data (ou sem data), mas não existia informação de identificação suficiente nos conjuntos de dados da HRVD e do RMS para os distinguir enquanto nomes que correspondiam a indivíduos distintos.

127. O procedimento de estabelecimento de correspondências P3 estabeleceu correspondências em que existiam iguais probabilidades de uma boa correspondência para um determinado registo, situação que não exigia a avaliação pelo especialista em ligação de registos.

Estabelecimento de correspondências entre violações fatais inter-sistemas aos pares

128. O estabelecimento de correspondências inter-sistemas para o par constituído pela HRVD e o RMS resultou numa lista nova de vítimas únicas, designada conjunto de dados HRVD/RMS. Este conjnto de dados incluía 10.594 registos existentes apenas no conjunto de dados HRVD, 4.087 existentes apenas no conjunto de dados RMS, e 532 que foram encontrados tanto na HRVD como no RMS. Estes 15.213 registos únicos foram depois feitos corresponder inter-sistemas com os 149.267 do conjunto de dados GCD, sendo que o conjunto de dados HRVD/RMS representava os dados-fonte e o conjunto de dados GCD os dados-alvo. O estabelecimento de correspondências aos pares entre o conjunto de dados HRVD/RMS e a GCD resultou em 157.000 nomeações de pessoas falecidas. Este total inclui registos que estavam fora do período de referência da Comissão ou não possuíam datas de falecimento que permitissem verificar se as pessoas em questão tinham morrido dentro do período de referência. Só foram usados para a análise os registos que possuíam datas de falecimento dentro do período de referência da Comissão.

129. As ligações no interior e entre estes conjuntos de dados são usadas para estimar o número total de mortos devidos ao conflito. Os registos nesta lista final podem ser ligados de volta a um único conjunto de dados, ou a uma combinação dos três conjuntos de dados. Apresenta-se a seguir uma matriz simples que mostra os resultados do procedimento final de estabelecimento de correspondências entre violações fatais inter-sistemas entre os conjuntos de dados.[*]

	Só HRVD	Só RMS	Só GCD	HRVD e RMS	HRVD e GCD	RMS e GCD	HRVD/ RMS/GCD	Total
Nº	5.203	2.148	141.787	382	5.391	1.939	150	157. 000
%	3,31	1,37	90,31	0,24	3,43	1,24	0,1	100

[*] Trata-se de totais não ponderados, e que incluem registos onde faltam datas, ou que apresentam datas que estão fora do período de referência, ou em que faltam indicações de local, ou que apresentam locais fora de Timor-Leste. Os registos que estavam fora do período de referência foram subsequentemente eliminados da análise.

130. Se o procedimento de estabelecimento de correspondências intra-sistema tivesse apanhado todas as correspondências possíveis, durante o estabelecimento de correspondências inter-sistemas só teria sido possível obter valores de correspondência potencial iguais a zero ou um. É possível escaparem correspondências se os registos que estão a ser examinados possuírem campos onde faltam os dados e que tornam difícil determinar se os dois registos deviam ter sido ligados. Os erros humanos também são possíveis, tendo em conta o grande volume de dados que o trabalho da Comissão envolveu. De um modo geral, uma correspondência é assumida quando uma maioria dos campos de dados se correspondem, ou quando a probabilidade de correspondência do registo está dentro do intervalo de tolerância. Quando não existe um número suficiente de campos com dados completos, torna-se difícil determinar com uma certeza razoável se um registo deve ou não ser excluído da possibilidade de ser feito corresponder com outro. Esta situação foi particularmente notória no caso de nomes indígenas timorenses muito comuns, como Mau Bere, quando havia muitas pessoas do mesmo local que tinham morrido ou sido mortas na mesma altura.

131. Depois de concluído o procedimento de estabelecimento de correspondências inter-sistemas em Analzyer, os dados foram importados para uma folha de cálculo para revisão pelo especialista em ligação de dados. Analisando os dados ordenados segundo diferentes variáveis, usando processos múltiplos —tanto humanos como automatizados—pode-se concluir com confiança que todas as correspondências que deveriam ter sido estabelecidas foram processadas. Adicionalmente, o procedimento de estabelecimento de correspondências inter-sistemas pode ser considerado uma medida da fiabilidade interavaliadores (*Inter-Rater Reliability*, *IRR*) na medida em que detecta instâncias em que na fase intra-sistema escaparam correspondências. Regressando aos dados intra-sistema e aplicando as correspondências que tinham escapado, foi possível não apenas medir a IRR mas também corrigir os dados, produzindo dados mais fiáveis sobre os quais basear as estimativas.

Table 3 - Tabela 3 - Nº totais e percentagens de registos em correspondência inter-sistemas para violações fatais por par de conjuntos de dados

Passo	HRVD e RMS	HRVD/RMS e GCD
Nº à partida	HRVD + RMS=HRVD/RMS	
Estabelecimento de correspondências na folha de	Número e Percentagem	
Ajustamento tendo em conta correspondências que	Número e Percentagem	
Total HRVD/RMS	Número e Percentagem	
Nº à partida		HRVD/RMS + GCD = MSE
Estabelecimento de correspondências P1		Número e Percentagem
Estabelecimento de correspondências P2		Número e Percentagem
Estabelecimento de correspondências P3		Número e Percentagem
Nº total para MSE		Número e Percentagem

6. Processamento de dados relativos a violações reportadas envolvendo grupos de vítimas anónimas

132. Durante o processo de recolha de testemunhos, um depoente pode ter-se referido a uma ou mais vítimas. Por vezes, quando um depoente se referia a múltiplas vítimas, o depoente não conhecia o nome de algumas ou de todas as vítimas. No processo de recolha de testemunhos da Comissão, em 1,9% (1.419/75.443) dos registos de vítimas documentados pela Comissão, o depoente não conhecia os nomes individuais das vítimas que tinham sofrido abusos no contexto de um grupo mais vasto de pessoas.

133. Para integrar esses dados na análise feita pela Comissão, e desse modo considerar não apenas as violações contra indivíduos nomeados mas também contra grupos anónimos, houve

necessidade de um processamento adicional dos dados para ter em conta prováveis registos duplicados de violações contra grupos de vítimas reportados. Os passos desse processamento para controlar tais duplicações

- identificaram registos de violações (contra vítimas anónimas de grupos) que pareciam descrever o mesmo grupo de vítimas, e depois
- escolheram um registo de vítima do conjunto de possíveis registos duplicados para ser considerado o "rep rec" dessa violação de vítima reportada.

134. Ao contrário dos dados sobre violações contra indivíduos (que, em grande medida, incluem identificadores pessoais tais como nomes, idades e sexo), as violações reportadas contra grupos não contêm geralmente identificadores do grupo-vítima. Por conseguinte, os registos de vítimas-grupo foram postos em correspondência através duma comparação das seguintes variáveis relativamente a cada violação reportada contra o grupo:

- o distrito onde a violação alegadamente ocorreu
- o tipo de violação em que a violação foi codificada, e
- o ano e mês em que a violação alegadamente ocorreu.

135. Depois de todos os registos grupo-vítima semelhantes serem postos em correspondência para formar um agregado, o registo com a maior dimensão de grupo dentro de cada agregado foi conservado. Todos os outros registos foram considerados registos duplicados e por iso eliminados do conjunto de dados.

136. O nível de duplicação entre registos de grupo-vítimas é apresentado na Figura {X}. Este quadro mostra quantos exemplares duplicados de violações foram identificados por tipo de violação no conjunto de dados e o número excedente de registos de violações em grupos que foram eliminados da análise da Comissão sobre violações contra vítimas de grupos.

	Prisão		Tortura		Maus-tratos		Deslocação		Outras violações		Todas violações	
Exemplares	Obs	Excesso	Obs	Excesso	Obs	Excesso	Obs	Excesso	Obs	Excesso	Obs	Excesso
1	441	0	134	0	121	0	180	0	736	0	1.612	0
2	150	75	26	13	30	15	68	34	206	103	480	240
3	69	46	15	10	9	6	21	14	87	58	201	134
4	56	42	4	3	8	6	16	12	60	45	144	108
5	25	20	0	0	5	4	10	8	30	24	70	56
6	6	5	0	0	6	5	12	10	12	10	36	30
7	0	0	0	0	7	6	0	0	0	0	7	6
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	12	11	0	0	0	0	0	0	12	11	24	22
13	13	12	0	0	0	0	0	0	13	12	26	24
Total	772	211	179	26	186	42	307	78	1.156	263	2.600	620

7. Técnicas de estimação estatística usadas na análise das violações fatais e deslocações

137. Nesta secção são apresentadas as técnicas de estimação baseadas nos estudos e os métodos de estimação de sistemas múltiplos (*multiple systems estimation*, MSE) usados para

produzir as estimativas da extensão total e padrão da mortalidade e deslocações durante o período de referência da Comissão.

Cálculos das ponderações no RMS

138. O modo como foi obtida a amostra do estudo foi atrás descrito: em 2003, as equipas da CAVR no terreno entrevistaram 1396 agregados familiares seleccionados de 138 aldeias e grupos de aldeias, designados agregados de aldeias (*clusters*). Os agregados de aldeias foram seleccionados através dum método designado "probabilidade proporcional à dimensão" (*Probability Proportional to Size*, *PPS*), e depois dez (ou 20) agregados familiares foram seleccionados através duma simples amostragem aleatória em cada um dos agregados de aldeias. Se cada agregado de aldeias tivesse exactamente o mesmo número de agregados familiares, a probabilidade de amostragem de cada agregado familiar seria idêntica, um processo conhecido como "autoponderação" (*self-weighting*).[8] Devido à recolha de amostras de 20 agregados familiares em agregados multi-aldeias e à ausência de respostas noutros agregados de aldeias, nem todos os agregados de aldeias possuíam igual número de agregados familiares; no entanto, 78.5% dos agregados de aldeias incluídos na amostra possuem exactamente 10 agregados familiares. As não respostas corresponderam a 3,1%, e por isso não foi introduzido qualquer ajustamento para não respostas. As ponderações foram calculadas como se segue.

139. Para cada agregado de aldeias (*cluster*), o ajustamento para a variação da dimensão do agregado é:

- cluster_adjustment = median_cluster_size / cluster_size

140. A probabilidade bruta de amostragem dos agregados familiares em 1990 é

- *sp_1990* = (nº total agregados familiares na amostra) / (total agregados familiares em 1990) = 1.396/168.858

141. E assim, para cada agregado de aldeias, a ponderação pps é

- pps_wt_1990_raw = (1/sp_1990) * cluster_adjustment

142. Entre 1990 e 2004, altura em que o estudo foi realizado, houve alterações significativas na população devido às migrações e ao seu crescimento. Antes das ponderações poderem ser estimadas, o número total de agregados familiares em cada aldeia foi ajustado em relação aos dados do censo de 1990 usando dados do censo de 2004. Quando a amostra estava a ser concebida, os agregados de aldeias foram escolhidos usando o número de agregados familiares para cada aldeia referidos no censo de 1990. Quando estes cálculos foram feitos (Abril 2005), os dados do Censo Timor-Leste 2004 estavam disponíveis desagregados apenas até ao nível do subdistrito, mas não do suco ou aldeia.[*] Note-se que os ajustamentos das ponderações 1990/2004 não afectam a ponderação somada total, que foi fixada de acordo com o número de agregados familiares que existiam 2004. Os ajustamentos das ponderações afectam o modo como os agregados em diferentes locais afectam as projecções.

143. Dois subdistritos listados no censo de 1990 não foram listados no censo de 2004: Fatumaca em Baucau foi absorvido pelo subdistrito de Baucau, e em Oecusse, Pante Macassar B foi subsumido em Pante Macassar. Para estes subdistritos, o número de agregados familiares em 2004 foi estimado usando a percentagem de agregados familiares nos subdistritos absorvente e absorvido em 1990 e multiplicando-a pelo total no subdistrito absorvente em 2004.

[*] Ver http://dne.mopf.gov.tp para os dados do censo.

144. Embora os números totais relativos aos agregados familiares em 2004 estejam disponíveis nos dados do censo a nível subdistrital, o RMS possui um número demasiado escasso de respostas a nível subdistrital para que as estimativas das ponderações por subdistrito possam ser feitas com dados adequados (29 dos 59 subdistritos da amostra apresentam menos de 20 respostas). Assim sendo, as ponderações de 1990 foram ajustadas aos totais distritais de 2004 através do seguinte cálculo:

- *district_adjustment* = (total agregados familiares em 2004 neste distrito)/(Total ponderação1990 neste distrito)*pps_wt_2004 = pps_wt_1990_raw * district_adjustment*

145. Forçando as ponderações a corresponderem ao número de agregados familiares por distrito no censo de 2004, as ponderações foram normalizadas de modo a que o seu somatório fosse igual ao número total de agregados familiares em 2004 (194.943). Os erros nos resultados foram calculados usando os *Stata standard survey modules*.[9] Estes módulos utilizam as variáveis na concepção do estudo (estrato, unidades de amostragem primárias e ponderação da amostra) para produzirem estimativas ponderadas dos totais e aproximações por séries de Taylor dos erros na amostragem. As estimativas de erro assumem uma amostragem aleatória com ponderações das amostras desiguais. Trata-se de um pressuposto conservador (ie. que tenderá a subestimar o erro de amostragem) relativamente às ponderações calculadas usando os métodos PPS atrás descritos[10] Os ficheiros de dados usados para estes cálculos estão disponíveis em http://www.hrdag.org/timor.

Atribuição de datas para análise de deslocações no RMS

146. O estudo perguntou aos inquiridos em que momento se deslocaram de cada um dos locais durante o período 1974/1999. Quando os inquiridos não estavam certos acerca de uma data específica nas suas deslocações, identificavam frequentemente o ano dessa deslocação e o momento do ciclo agrícola, ou se se tratava da época seca ou das chuvas. Para cada uma destas datas parciais ou sazonais, atribuímos a deslocação ao trimestre que incluía o período ou estação. Nos casos em que a identificação parcial da data podia cair em mais de um trimestre, foi atribuída aleatoriamente a um deles. Dos 2.024 movimentos definidos pelos inquiridos como deslocações, 76,6% foram identificados pelos menos até ao trimestre, e outros 15,7% foram identificados pela estação. Apenas 7,7% das deslocações foram identificadas por ano sem qualquer especificação do mês.

Ajustamentos das ponderações para estimação da mortalidade no RMS

147. O cálculo das ponderações assume que os acontecimentos reportados por cada agregado familiar podiam apenas ter sido reportados por esse agregado familiar. Este pressuposto é o resultado das ponderações serem simplesmente o inverso da probabilidade de amostragem para um dado agregado familiar. Por conseguinte, existindo mais de um agregado familiar que pudesse ter fornecido informações acerca duma morte específica, a verdadeira probabilidade de amostragem para essa morte seria maior do que a probabilidade para um único agregado familiar. As mortes reportadas pelos inquiridos do estudo violam o pressuposto do único-agregado-familiar-que-reporta-o-acontecimento porque para cada morte, pode ter havido mais de um agregado familiar a fornecer informações sobre essa morte. De entre o total de 5.402 mortes reportadas pelos inquiridos, 545 foram reportadas mais de uma vez (os relatos duplicados foram identificados e removidos antes da estimação). Os reportes duplicados implícitos na ponderação do estudo foram corrigidos através dum ajustamento das ponderações do modo que se descreve adiante.

148. Antes das ponderações do estudo poderem ser usadas para estimar o número total de mortes, têm de ser ajustadas para dar contado número de agregados familiares que potencialmente forneceram informações sobre cada morte. Ou seja, para cada morte, quantos familiares sobreviveram até 2003 para serem inquiridos potenciais do estudo? Muita da

informação necessária para este cálculo está disponível no estudo porque os familiares do inquirido são também os familiares do falecido. O número de familiares sobrevivos para cada falecido (*decedent*) *D* foi calculado com base nos familiares reportados pelo inquirido (*respondent*) *R* usando as seguintes regras:

1. Se *D* é um progenitor de *R*, o número esperado de familiares sobrevivos em 2003 é a soma do seguinte*:*

- Assumir que os progenitores de *D* são 25 anos mais velhos do que *D* (ou 50 anos mais velhos do que *R*, se a idade de *D* não for reportada); usar probabilidades condicionais de sobrevivência específicas para a idade (calculadas a partir do estudo) para estimar o número esperado de progenitores vivos em 2003
- Contar os irmãos de *R* como filhos de *D*
- Considerando uma taxa de fertilidade total aproximada média de 5 antes de 1975, assumir que *D* tem quatro irmãos com idades de (-4, -2, +2, +4) anos do que a idade de *D* (se a idade de *D* for desconhecida, considerar a idade de *D* igual à de *R* + 25), calcular as idades dos irmãos em 2003, e multiplicar cada uma pela probabilidade condicional de sobrevivência até essa idade, e somar os valores para os quatro irmãos.

2. Se *D* é irmão de *R*

- Os progenitores de *D* são os progenitores de *R*, contar os sobrevivos directamente
- Os irmãos de *R* são os irmãos de *D*, contar os sobrevivos directamente

149. Assumir que D tem o mesmo número de filhos adultos sobrevivos que R.

3. Se *D* é filho de *R*

- *R* e cônjuge são progenitores, contar os sobrevivos directamente
- Os filhos adultos de *R* são irmãos de *D*, contar os sobrevivos directamente
- Assumir que não existem quaisquer filhos adultos de *D* sobrevivos.

150. Este cálculo produz o número esperado de familiares adultos sobrevivos para cada *D*, para além de indicar quais destes familiares sobrevivos vivem no agregado familiar de *R*, e quais vivem noutros agregados familiares.

151. Para converter os familiares adultos sobrevivos esperados de *D* num factor de ajustamento para a ponderação da amostra, o número de familiares deve ser convertido num número esperado de agregados familiares em que os familiares vivem. Existem em média 0,5 familiares de *D* (para além de *R*) vivendo no agregado familiar de *R*. Assumir que os outros agregados familiares onde vivem familiares de *D* apresentam a mesma concentração de familiares por agregado que o agregado familiar de *R* (ou seja, 1,5 familiares por agregado). Assim, se *D* tiver *L* familiares sobrevivos que vivem for a do agregado familiar de *R*, existem $a = 1 + L/1{,}5$ agregados que poderiam dar informações sobre *D*. As ponderações do estudo foram ajustadas para ter em conta a possibilidade de reportes múltiplos relativamente *D,* dividindo a ponderação da amostragem para cada *D* por este factor, *a*. Este cálculo assume que os outros agregados familiares que potencialmente podem dar informações estão no mesmo agregado de *R*, ou que estão num agregado com uma probabilidade de amostragem no interior do agregado que é semelhante.

Análise de sensibilidade dos pressupostos em reponderações de mortalidade

152. Existem diversos pressupostos nos ajustamentos das ponderações para as estimativas de mortalidade, incluindo os seguintes:

- A diferença de tempo entre gerações (que se assume ser de 25 anos)
- O número de irmãos que os progenitores inquiridos tinham (que se assume serem quatro)
- O espaçamento entre os nascimentos dos irmãos dos progenitores (que se assume ser de dois anos)
- O número de filhos adultos que os irmãos da pessoa inquirida tiveram (que se assume ser igual ao de filhos do inquirido).

153. Estes pressupostos foram testados usando as seguintes variações, e o número total anual de mortes foi calculado:

- O espaçamento entre gerações foi feito variar entre 18 e 30 anos
- Considerou-se que o número de irmãos dos progenitores do inquirido aumentara para seis
- O espaçamento dos nascimentos foi aumentado para cinco anos entre irmãos
- O número de filhos adultos de irmãos do inquirido foi considerado o dobro do número de filhos do inquirido.

154. Para cada estimação duma variante, os totais anuais foram testados (através de um teste-t de diferença de médias) em relação ao modelo principal. Nenhum dos anos em qualquer dos modelos variantes era significativamente diferente a $p<0,05$. O valor mínimo de p foi 0,13, e estava à margem dos restantes valores (*outlier*): o segundo valor de p mais baixo foi 0,23. Por conseguinte, as estimativas não são substancialmente sensíveis aos pressupostos acerca da estrutura familiar.

155. Embora as estimativas sejam robustas relativamente aos pressupostos sobre as estruturas familiares usados para estimar o número de familiares sobrevivos que poderiam dar informações sobre D, as magnitudes das estimativas são sensíveis ao modelo utilizado para transformar o número estimado de familiares sobrevivos no número estimado de agregados que contêm familiares. O número estimado de familiares sobrevivos é L, e o número estimado de agregados que contêm familiares de um falecido D, designado a, é $a = 1 + L/1,5$. O denominador 1,5 resulta do número médio de familiares de D (incluindo R) que vivem no agregado familiar de R (0,5). Fazendo variar esta média de 0 a 3 (ou seja, assumindo 1-4 familiares adultos sobrevivos por agregado familiar), as estimativas do número total estimado de mortes (todas as causas) resultantes variam entre –14,2% to +19,6%. O efeito das variações neste modelo descresce ao longo do tempo, situando-se as variações mais significativas nos anos iniciais 1972/1975 (-21%, +26%) e as variações mais pequenas em anos mais recentes 2001/2003 (-11%, +16,2%). Este decréscimo é consistente ao longo do tempo.

156. Considerando um número constante de familiares sobrevivos, menos familiares sobrevivos por agregado familiar significam um maior número de agregados familiares que podem potencialmente dar informações, uma probabilidade de amostragem estimada mais elevada por morte reportada, e uma ponderação da amostra mais baixa por morte reportada e, por conseguinte, um número total de mortes estimadas menor; um maior número de adultos por agregado familiar inverte esta lógica.

157. Embora as estimativas totais variem com as mudanças no modelo que transforma familiares em agregados familiares, os padrões são constantes. Os coeficientes de correlação do modelo principal com os modelos baixo (0) e alto (3) acima referidos são cada qual de 0,99. Embora o modelo de familiares-por-agregado familiar afecte a magnitude total do número de mortes estimado, não afecta os padrões estimados ao longo do tempo.

Estimação de sistemas múltiplos (*Multiple Systems Estimation*, MSE): motivação e teoria

158. A análise do estudo é conservadora no sentido em que corrige potenciais reportes em duplicado através do estabelecimento de correspondências entre mortes nos agregados familiares, e porque existe um ajustamento às ponderações da amostragem baseado no número estimado de agregados familiares que poderiam ter dado informações acerca de cada morte. Uma vez que algumas mortes podem ser reportadas por vários agregados familiares, existem outras mortes que ocorreram em 1974/1999 e para as quais não existiam familiares sobrevivos em 2003. Se agregados familiares inteiros tiverem morrido durante o período de referência da Comissão, não haverá em 2003 familiares colineares para fornecerem informações. Tendo em conta estas limitações, um método alternativo para estimar o número total de mortes pode constituir um mecanismo de controlo das estimativas do estudo.[*]

159. O método MSE utiliza diversas listas incompletas da população recolhidas separadamente. As listas são postas em correspondência para identificação dos elementos comuns nas várias listas, a fim de se estimar o número de elementos que faltam em todas as listas. Neste projecto, as mortes documentadas na HRVD, RMS, e GCD foram postas em correspondência nos três sistemas, usando o nome, data de falecimento, local de falecimento e data de nascimento.

160. A forma mais básica desta técnica é designada captura-etiquetagem-recaptura (*capture-tag-recapture*), e recorre apenas a duas listas.

161. Uma explicação técnica sobre o modo como o número dos membros desconhecidos duma população pode ser estimado é apresentada de seguida. Considere-se o caso de dois projectos P_1 (uma lista de *A* indivíduos) e P_2 (uma lista de *B* indivíduos). Existem *M* indivíduos que são feitos corresponder nas duas listas, num universo total de *N* indivíduos (*N* é desconhecido). Se todas as pessoas no universo *N* tiverem igual probabilidade de aparecer na Lista 1, a probabilidade de um indivíduo específico ser reportado pelo P_1 é dada por

162. *Pr(capturado na lista 1)* = $\frac{A}{N}$

163. Analogamente, se todas as pessoa no universo *N* tiverem igual probabilidade de aparecer na Lista 2, a probabilidade de um indivíduo específico ser reportado pelo P_2 é dada por

164. *Pr(capturado na lista 2)* = $\frac{B}{N}$

165. A probabilidade dum indivíduo específico ser capturado em ambas as listas é dada por

Pr(capturado na lista 1 e na lista 2) = $\frac{M}{N}$

[*] Esta explicação segue P Ball, J Asher, D Sulmont, D Manrique, "How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000", a report to the Peruvian Truth and Reconciliation Commission, Washington, DC: AAAS, 28 August 2004. Disponível em http://shr.aaas.org/hrdag/peru

166. Por definição, a probabilidade dum acontecimento composto por dois acontecimentos independentes é o produto das duas probabilidades independentes. Por isso,

Pr(capturado nas listas 1 e 2) = Pr(capturado na lista 1) x Pr(capturado na lista 2)

167. Que é $\frac{M}{N}=(\frac{A}{N})\cdot(\frac{B}{N})$: dada esta equação, resolver para *N.* Rearranjando os termos, $\frac{M}{N}=\frac{A\cdot B}{N^2}$ e multiplicando depois por *N*, $M=\frac{A\cdot B}{N}$ multiplicando de novo $M\cdot N=A\cdot B$, e finalmente dividindo por M dá $N=\frac{A\cdot B}{M}$. Note-se que com a equação final, o número total de mortes *N* pode ser estimado usando os totais de *A* e *B* e as correspondências entre eles, *M*.

168. Existem muitos pressupostos implícitos nesta solução. Por exemplo, que nenhuma das listas tem indivíduos reportados duas vezes e que o estabelecimento de correspondências entre as listas é rigoroso. Neste projecto, estes dois pressupostos foram controlados durante o processamento dos dados da forma descrita na secção sobre o estabelecimento de correspondências.

169. Existem outros pressupostos inerentes ao modelo captura-etiquetagem-recaptura que são mais difíceis de gerir. Em primeiro lugar, o método assume que os indivíduos não estão a entrar ou a sair do universo definido durante o processo de criação de listas. E, em segundo lugar, que as listas foram seleccionadas aleatoriamente na população. Nos projectos de documentação de direitos humanos, o primeiro pressuposto é geralmente irrelevante porque a documentação é feita retrospectivamente. O segundo pressuposto não pode ser satisfeito, e tem de ser substituído pelo pressuposto de que a estimação é robusta relativamente ao processo de selecção.

170. Um outro pressuposto é de que as listas são independentes, ou seja, de que a probabilidade de um indivíduo estar na lista dois é independente da probabilidade desse indivíduo ser capturado na lista um. O último pressuposto diz respeito à homogeneidade: trata-se de supor que os indivíduos que compõem o universo têm todos igual probabilidade de serem capturados.

171. Se qualquer um destes pressupostos for violado, o método de captura-etiquetagem-recaptura não produz uma estimativa adequada da dimensão total da população. Não existindo mais de duas listas com informação adequada, os problemas de dependência ou heterogeneidade podem frequentemente ser resolvidos através da especificação e selecção de modelos apropriados. Contudo, nos dados da HRVD, RMS, e GCD, só existem dois sistemas utilizáveis (RMS-GCD para mortes devidas a fome e doenças, e HRVD-GCD para os assassinatos).[*] Por si sós, estas estimativas seriam insuficientes, mas combinadas com as estimativas do RMS, fornecem úteis informações adicionais.

[*] A aplicação inicial de uma estimação de sistemas múltiplos a estimações de natureza demográfica deveu-se a C Chandra Sekar e W Edwards Deming, "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, Março 1949. Uma análise detalhada dos estimadores para o método do sistema dual e dos cálculos de erro relevantes pode ser encontrada *in* Yvonne M M Bishop, Stephen E Fienberg e Paul H. Holland. *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, 1975. Para um comentário acerca da utilização destes métodos em análisee de direitos humanos, ver Fritz Scheuren, "History Corner," *The American Statistician*, Fevereiro 2004.

Atribuição da GCD por tipo de morte

172. Os dados relativos aos cemitérios não incluem informações sobre o modo como as mortes ocorreram. Existiam 89.894 sepulturas com pelo menos uma primeira inicial (ou nome), um último nome e um ano de falecimento entre 1972 e 2003. Destes dados, 7.117 foram postos em correspondência com a HRVD ou o RMS (ou ambos), e através destas correspondências é possível determinar o modo como ocorreu a morte a partir do modo como ocorreu a morte nos registos correspondentes. Os restantes 82.717 registos da GCD tiveram de ser atribuídos às quatro categorias de modos de ocorrência da morte (assassinatos, mortes devidas a fome e doenças, mortes de combatentes, e outras mortes). As percentagens anuais de mortes destes quatro tipos no RMS são apresentadas na Figura <number>, adiante. Note-se que estas percentagens excluem as mortes em que o modo de ocorrência da morte é desconhecido (204 de 3.235 mortes reportadas no RMS entre 1969 e 2004 têm um modo de ocorrência da morte desconhecido).

Table 4 - Figura <number>: percentagens estimadas de mortes, por período e modo de ocorrência da morte

Período	Morte ilícita	Fome/Doença	Combatente	Outro
1972/1974	0,9%	95,9%	0,0%	3,2%
Margem de erro	1,8%	5,1%	0,0%	4,9%
1975/1982	11,2%	83,0%	4,4%	1,4%
Margem de erro	4,7%	5,1%	2,5%	0,6%
1983/1998	5,5%	86,5%	0,7%	7,2%
Margem de erro	2,5%	3,7%	0,6%	2,5%
1999	16,2%	83,0%	0,4%	0,4%
Margem de erro	10,2%	10,2%	0,8%	0,8%
2000/2003	3,5%	86,9%	0,8%	8,9%
Margem de erro	3,1%	6,5%	1,6%	4,9%
Total	8,3%	85,1%	2,4%	4,3%
Margem de erro	2,7%	31%	1,2%	1,2%

173. Estas percentagens foram usadas para atribuir os registos da GCD para os quais não tinham sido encontradas correspondências aos diferentes modos de ocorrência da morte para serem usados nos cálculos MSE para cada ano: as percentagens do período contendo cada ano foram usadas para atribuir as mortes da GCD nesse ano. A margem de erro da atribuição foi incluída no erro estimado das estimativas obtidas pelo método MSE.

Análise de sensibilidade da perda de conhecimento social: ajustamentos para as situações de subestimação

174. O estudo interrogou os inquiridos acerca das mortes dos seus progenitores, irmãos e filhos. No entanto, algumas mortes não deixaram quaisquer progenitores, irmãos ou filhos que ainda estivessem vivos quando o estudo foi realizado em 2004. Se as mortes tivessem ocorrido no passado longínquo, mesmo os filhos dos falecidos teriam todos morrido, não deixando ninguém para reportar essas mortes. Noutros casos, famílias pequenas podiam ter sofrido mortalidade completa, não tendo sobrevivido ninguém para reportar essas mortes. À medida que o estudo estima o número (ou a taxa) de mortes para períodos cada vez mais distantes no

passado, a subestimação que resulta da perda de conhecimento social tende a acentuar-se. Contudo, mesmo para o passado mais imediato (por exemplo, em 2003 para um estudo realizado em 2004), será impossível documentar algumas mortes que não deixaram sobreviventes. Por exemplo, as pessoas que não deixaram progenitores, irmãos ou filhos sobrevivos e que morreram em 2003 não podem ser referidas no estudo.

175. A taxa de mortalidade bruta (por 1000 pessoas) é uma estimativa do número de pessoas que morreram, no total, em cada ano. Trata-se dum indicador demográfico e de saúde habitual, que é geralmente estimado por métodos indirectos utilizando os dados dos censos. No caso de Timor-Leste, essas taxas são difíceis de estimar porque a qualidade dos dados dos censos de 1980 e 1990 tem sido questionada.[11] Apresentam-se as taxas de mortalidade bruta (*Crude Death Rate*, CDR) estimadas pelo US Bureau of the Census para Timor-Leste para 1990/2004. A taxa global indonésia é indicada para 1983. A estimativa apresentada para 1971 resulta duma afirmação do Governo indonésio de que, em toda a Indonésia, a CDR diminuiu 45% entre 1971 e 1990; a estimativa para 1971 aqui apresentada é a estimativa de 1990 para Timor-Leste a que foi aplicado esse factor. Uma CDR projectada obtida através da interpolação linear entre a estimativa de 1971 e as estimativas de 1990/2004 é também apresentada.

176. Para além das estimativas da CDR, apresenta-se a CDR do RMS realizado pela Comissão. Esta estimativa representa o número total estimado de mortes dividido pela população estimada para esse ano (multiplicado por 1000). Há várias observações a fazer acerca deste gráfico. Em primeiro lugar, a CDR estimada pelo US Census Bureau situa-se no intervalo de confiança da CDR estimada a partir do RMS a partir de 1993. Em 2003, o intervalo de confiança da CDR do RMS (4,2 – 6,6) inclui a estimativa do US Census Bureau (6,4), como se mostra no gráfico pelo pico no final da linha da CAVR. Ou seja, enquano que o RMS subestima significativamente a taxa de mortalidade nos anos de paz “normais” entre 1972/1974, em meados da década de 80 o RMS está em conformidade com os resultados obtidos através de métodos indirectos pelo US Census Bureau. Esta observação é consistente com a ideia de que as estimativas obtidas a partir do RMS sofrem de uma subestimação crescente à medida que se caminha para o passado.

177. Durante os anos para os quais os registos históricos sugerem ter existido um número substancial de mortes em excesso, a interpolação linear da CDR subestima o número de mortes. Estes anos incluem 1975/1979 e 1999. Isto é consistente com o sentido literal da expressão “mortes em excesso”. (Não existem estimativas da CDR baseadas nos censos para o período 1975/1979). Olhando para o passado mais longínquo, a CDR baseada no estudo captura uma proporção decrescente da CDR total (é possível desenhar um gráfico análogo para as estimativas ao longo do tempo obtidas pelo método MSE, com resultados semelhantes).

178. Para ajustar o RMS, as mortes perdidas devida à perda do conhecimento social têm de ser estimadas ao longo do tempo. O modelo utilizado foi o seguinte:

- o número de mortes estimado pela CDR e a população projectada para cada ano foram estimados (“mortes CDR”), sendo os valores apresentado sob a forma de uma taxa na Figure {g_cdrs.pdf};
- a fracção das “mortes CDR” que se deveram a fome e doenças foi estimada usando a fracção de todas as mortes reportadas no estudo que se deveram a fome e doenças (de modo análogo ao que foi usado para o processo de atribuição dos dados da GCD para os quais não tinham sido encontradas correspondências). No estudo, a fracção média (e mediana) de todas as mortes (ao longo dos anos) que se deveram a fome e doenças é de 0,80, e 50% de todos os anos estão no intervalo 0,754 – 0,846;
- o rácio de mortes estimadas para “mortes CDR” foi calculado para os anos de paz (1972/1974 e 2002/2003); esta é a fracção de “mortes rememoráveis,” designada “fracção de memória” (*memory fraction*);
- A fracção de memória para 1975/2001 foi estimada por interpolação linear usando a seguintes equações:
- fracção de memória estimada (MSE) = -39,1 + 0,0200*ano
- fracção de memória estimada (RMS) = -43,9 + 0.0224* ano

- As fracções de memória para intervalos MSE situam-se entre 0,241-0,936, enquanto que para o RMS, se situam entre 0,228 e 0.846. Esta diferença tem um impacte enorme sobre o resultado final.
- A estimativa ajustada foi calculada como sendo igual à estimativa original dividida pela fracção de memória para cada ano.

179. As estimativas ajustadas são apresentadas a seguir, nas Figuras {g_huil_xs_mse.pdf} e {g_huil_xs_rms.pdf}. Note-se que em ambos os gráficos as estimativas iniciais e as estimativas ajustadas convergem à medida que o ano se aproxima de 2003. O impacte da fracção de memória maior nos valores obtidos pelo método MSE relativamente aos valores obtidos do RMS é visível no número total estimado de mortes que excedem o valor de referência da CDR: a estimativa ajustada obtida pelo método MSE é de 104.000 mortes enquanto que a estimativa ajustada obtida a partir do RMS é de 183.300 mortes.

180. Ambas as estimativas dependem dum conjunto de pressupostos, incluindo pressupostos acerca da forma do decréscimo da CDR a partir do início da década de 70 e até ao final da década de 90, bem como acerca da natureza da perda de memória social. As alterações suaves mas não lineares na perda de memória social (concavidade para baixo ou para cima) não alterariam substancialmente a estimativa. No entanto, se a subestimação nos valores obtidos por MSE e do RMS devida a perda de memória social fosse de algum modo descontínua ou drasticamente diferente para o período 1972/1974 relativamente aos anos de pico 1975/1979, o ajustamento aqui utilizado não seria capaz de corrigir adequadamente essa subestimação. Ambos os modelos dependem de CDRs calculadas a partir dos censos de 1980 e 1990 e dos métodos indirectos usados pelo US Bureau of the Census. Existe erro de amostragem e não amostragem que não está representado nos gráficos ou nas estatísticas, mas que é certamente substancial.

181. No entanto, estes modelos têm a vantagem de mostrarem que, com o ajustamento, as mortes totais anuais estimadas devidas a fome e doenças se aproximam das “mortes CDR” de referência para o período pré-invasão (1972/1974) e para o período 1984/1998.

182. Existem várias razões para preferir a estimativa obtida pelo método MSE à estimativa obtida a partir do RMS. Embora o valor do RMS esteja mais próximo da estimativa de mortes CDR nos anos pós-ocupação que se aproximam do tempo de paz, 2002/2003, o valor obtido por MSE está mais próximo das estimativas de mortes CDR totais no período pré-ocupação. Para

efeitos desta estimativa, o período mais relevante é 1975/1979, e a escolha da estimativa deve guiar-se pelos valores que melhor se adequam imediatamente antes do início desse período. Uma Segunda razão para preferir o valor obtido pelo método MSE é o facto de se basear num volume consideravelmente maior de dados do que o valor obtido do RMS: o valor obtido pelo método MSE recorre aos dados da GCD para além dos dados do RMS.

183. A conclusão mais forte que pode ser extraída é de que as estimativas não ajustadas obtidas do RMS e pelo método MSE devem ser demasiado baixas. Na Parte 6: Perfil das Violações de Direitos Humanos, é apresentada uma análise da fundamentação estatística que suporta as conclusões relativas ao número de violações fatais ocorridas durante o período de referência da Comissão.

---

[1] UNTAET, Regulamento nº 2001/10, artº 13, nº 1º a) (i).

[2] UNTAET, Regulamento nº 2001/10, artº 13, nº 1 a) (i).

[3] UNTAET, Regulamento nº 2001/10, artº 13, nº 1 a) (i).

[4] UNTAET, Regulamento nº 2001/10, artº 13, nº 1 a) (ii).

[5] UNTAET, Regulamento nº 2001/10, artº 13, nº 1 a) (iv).

[6] UNTAET, Regulamento nº 2001/10, artº 13, nº 1 d).

[7] Patrick Ball, *Who Did What to Whom Handbook*, e Patrick Ball et al, *HR Database Design Methods*. US.

[8] Paul S Levy and Stanley Lemeshow, Sampling of Populations, Chapter 11, Wiley, New York, 1999.

[9] Stata Corporation, Stata Survey Data Reference Manual, v. 8, College Station, TX: Stata, 2003.

[10] Donna Brogan, "Sampling error estimation for survey data," *in* Household Sample Surveys in Developing and Transition Countries, United Nations Publication ST/ESA/STAT/SER.F/96, Department of Economic and Social Affairs of the United Nations Secretariat, 2005.

[11] See for example, Ben Kiernan, "The Demography of Genocide in Southeast Asia: The Death Tolls in Cambodia, 1975-79, and Timor-Leste, 1975-80." *Critical Asian Studies* 35:4 (2003), pp. 585-597.