# 7
# Considerações finais

Nos últimos anos, a integração de dados tem recebido cada vez mais atenção, devido ao crescente número de fonte de dados disponíveis e à necessidade de estudos da relação de fenômenos de diferentes tipos, que exigem abordagens justificadas de acordo com seus diferentes paradigmas.

Ressalte-se que a integração de dados é também um problema estatístico, além de representar um problema da tecnologia da informação. Inicia-se esse trabalho dizendo que o emparelhamento estatístico é uma ferramenta rápida, de baixo custo e flexível para integrar dados de diferentes fontes. Entretanto, alguns problemas podem surgir, e esses necessitariam de estudos mais profundos para que a aplicação de emparelhar estatisticamente seja bem sucedida.

O primeiro deles, que antecede o emparelhamento estatístico, apesar de não ter sido usado nesse estudo, é o processo para harmonizar duas diferentes fontes de dados, pode revelar-se de alto custo em tempo e dinheiro.

O segundo problema é a especificação do modelo de acordo com a informação disponível. Como já mencionado, caso exista informação auxiliar disponível, ela deveria sempre ser usada no processo de emparelhamento, uma vez validada a sua adequação. Sejam as abordagens micro ou macro, a escolha da metodologia de emparelhamento estatístico depende da combinação entre o objetivo do processo de integração e da informação disponível. E essa escolha deve delinear a qualidade resultante do ato de emparelhar.

O terceiro grande problema é a avaliação da qualidade, que indica que a escolha do procedimento também deveria depender da qualidade da entrada de

dados preservada na saída dos mesmos. Uma série de métodos deveria ser desenvolvida para estimar o viés e a variância, de cada estimador que tenha sido aplicado para concatenar o arquivo. Na prática, um artigo de Ingram et. al. (2000) faz uma avaliação, comparando as variáveis conjuntamente coletadas na pesquisa NSAF – *National Survey of American Families*, que eram obtidas a partir do emparelhamento estatístico, da área de saúde, que tem sido realizado, entre os arquivos NHIS – *National Health Interview Survey* - e o CPS – *Current Population Survey*.

Finalmente, deve-se lembrar que a maioria dos problemas que surgem quando se aplica o emparelhamento estatístico poderia ser evitado se utilizássemos questionários aninhados (Kroese et. al. , 2000), que é o uso da amostragem matricial (Rodrigues, 2003).

Mencionada na seção (1.4), o uso da amostragem matricial evitaria a harmonização de diferentes fontes, usando as mesmas definições para as variáveis comuns e a mesma metodologia para lidar com problemas usuais de pesquisa. Adicionalmente, seria possível melhorar a qualidade global dos dados observados desenhando questionários menores que então reduziriam o percentual de não resposta parcial e total. Ressalte-se que a redução da carga de resposta pode significar a obtenção de melhores taxas de respostas, o que passa a representar uma melhor qualidade da pesquisa, conseqüência de sua nova natureza amostral. Finalmente, pesquisas com questionários divididos aninhados, podem ser planejados com a finalidade de coletar a quantidade de informação necessária para uma estimativa razoável do modelo escolhido.

**Recomendação:**

A evidência disponível sugere que os resultados empíricos do estudo proposto confirmaram que o uso de informação auxiliar é uma vantagem quando a CIA não é válida. O grau de proteção depende do método e do tipo de informação auxiliar usada. Os procedimentos *hot deck* previamente descritos usaram alguns estimadores não-paramétricos como o *kNN*, com k igual a 1, que é método de *distance hot deck*.

A lista de novas investigações a serem implementadas sugeridas são:

- procedimentos de concatenação que considerem os pesos amostrais;
- procedimentos de concatenação com abordagem restrita (*constrained*) ;
- ampliar a dimensão da variável $X$ de concatenação;
- investigar o uso de outros estimadores não-paramétricos
- investigar o *rank hot deck*