ARTIGO

Cruzamento de dados do SINASC e SIM para estudos sobre mortalidade infantil: uma contribuição para a gestão da saúde pública

Ivan Roberto Ferraz
Doutorando em Administração pela Universidade de São Paulo (USP), Mestre em Administração pela Pontifícia Universidade Católica de São Paulo (PUC/SP), Especialista em Análise de Dados e Data Mining pela Fundação Instituto de Administração (FIA) e Bacharel em Administração pela Universidade Presbiteriana Mackenzie.
ivanferraz@hotmail.com

Elza Okubo
Especialista em Análise de Dados e Data Mining pela Fundação Instituto de Administração (FIA) e Bacharel em Estatística pela Universidade Estadual de Campinas (UNICAMP)
elza.okubo@uol.com.br

Alessandra de Ávila Montini
Doutora em Administração de Empresas pela Universidade de São Paulo (USP), Mestra e Bacharel em Estatística pelo Instituto de Matemática e Estatística (IME-USP)
amontini@usp.br

RESUMO: Ao propor uma metodologia de cruzamento das bases de dados do Sistema de Informações sobre Nascidos Vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM) para estudos sobre mortalidade infantil, este estudo defende que bases de dados de domínio público podem ser utilizadas de maneiras inovadoras para gerar conhecimento útil ao gestor público. O potencial da metodologia sugerida é a exemplificado por meio de um estudo que utiliza a técnica de regressão logística para identificar os fatores individuais relacionados à mortalidade, no primeiro ano de vida, dos nascidos vivos no estado de São Paulo nos anos de 2006 e 2007. Os resultados revelam a anomalia congênita, índice de Apgar no 1º e 5º minuto e peso ao nascer como os fatores que apresentaram mais impacto na mortalidade infantil.

PALAVRAS-CHAVE: Mortalidade infantil; Bases de dados públicas; Regressão logística.

Merging data from SINASC and SIM for studies on infant mortality: a contribution to public health management

ABSTRACT: By proposing a methodology to merge data from the Brazilian databases of live births (SINASC) and mortality (SIM) for studies on infant mortality, this paper argues that public databases may be used in innovative ways to generate knowledge useful to public managers. The potential of the suggested methodology is exemplified by a study that uses the technique of logistic regression to identify the individual factors related to mortality in the first year of life, of all births in the state of São Paulo between 2006 and 2007. The results indicate that congenital abnormality, Apgar scores at the 1st and 5th minutes and birth weight are the factors that have the greatest impact on infant mortality.

KEY WORDS: Infant mortality; Public databases; Logistic regression.

1 Submetido em 20 de abril de 2012. Aceito em 7 de dezembro de 2012. O artigo foi avaliado segundo o processo de duplo anonimato além de ser avaliado pelo editor. Editores responsáveis: Márcio Augusto Gonçalves e Lucas Maia dos Santos.

1. INTRODUÇÃO

Foram muitos os avanços tecnológicos nas últimas décadas: *hardwares* com poder de processamento cada vez maior, *softwares* mais amigáveis, infraestruturas de rede que possibilitam comunicações instantâneas entre qualquer região do planeta, etc. Estes e tantos outros avanços permitem hoje gerar e armazenar quantidades de dados gigantescas a um custo acessível, algo que seria inconcebível há apenas algumas décadas.

O desafio atual é tratar, relacionar e analisar esse grande volume de dados e gerar informações relevantes para a tomada de decisão. Organizações como governos e empresas privadas têm investido muitos recursos com o intuito de aproveitar todo o potencial disponível em bases de dados acumuladas durante anos, mas pouco aproveitadas até então.

Nesse sentido, existem muitas bases de dados de domínio público disponíveis na internet, das quais é possível extrair conhecimento útil para aperfeiçoamento de políticas públicas de educação, saúde, segurança, etc. Quanto mais alto o número de pesquisadores trabalhando nessas bases, mais chances há de que a gestão municipal, estadual ou federal se beneficiem de todo o conhecimento nelas existente.

Na área de saúde, por exemplo, por meio do Sistema de Informações sobre Nascidos Vivos (SINASC) e Sistema de Informações sobre Mortalidade (SIM), são disponibilizadas, respectivamente, bases de dados com registros de nascidos vivos e registros de óbitos em todo o território nacional. Registros de vários anos podem ser utilizados por qualquer pesquisador interessado, uma vez que tais bases estão disponíveis na Internet no *site* do Departamento de Informática do SUS (Ministério da Saúde).

De fato, muitos pesquisadores usam essas bases em suas pesquisas. Uma simples busca no *Google* Acadêmico retorna milhares de resultados para os termos SINASC e SIM. Entretanto, seria possível encontrar uma forma inovadora de utilizar essas bases para gerar informações novas e relevantes? A proposta deste artigo é discutir esta questão de modo a inspirar outros pesquisadores a pensar em formas inovadoras de trabalhar a imensa riqueza de dados secundários acumulados nas últimas décadas e que hoje são acessíveis ao pesquisador interessado.

No caso das bases do SINASC e SIM, são poucos os trabalhos que utilizam essas bases de forma integrada. O simples fato de relacionar os registros de nascimentos e óbitos para indivíduos que faleceram no primeiro ano de vida, por exemplo, seria uma nova forma de utilizar os mesmos dados e tal inovação metodológica poderia fornecer informações relevantes para o gestor de saúde preocupado com a questão da mortalidade infantil. Trabalhos como o de Morais Neto (1996) fazem essa integração empregando etapas manuais que envolvem pesquisas detalhadas em registros de hospitais, por exemplo. Esse tipo de procedimento, apesar de bastante preciso, é lento, custoso e inviável para volumes de dados muito grandes. Todavia, não haveria uma forma mais rápida e barata de integrar essas bases, ainda que menos precisa?

Cruzando os dados das duas bases, é possível analisar os fatores ou atributos individuais do recém-nascido vivo, da gestante e da gestação que têm alta associação com a mortalidade infantil. Isso ajuda a concluir quais são as variáveis que constituem fatores de risco de óbito antes de completar um ano de vida.

Estudos que possam contribuir para a redução da taxa de mortalidade infantil estão, mesmo que indiretamente, ajudando a salvar vidas. Muitos trabalhos e ações já foram realizados nesse sentido nos últimos anos. A taxa de mortalidade infantil no Brasil teve queda acentuada nas últimas décadas. Particularmente no estado de São Paulo, essa queda foi bastante forte na década de 90 e nos últimos anos o ritmo diminuiu, estabilizando o índice de mortes por mil habitantes em um patamar abaixo de 20, o que é considerado um bom índice (RIPSA, 2002). Todavia, a situação não é homogênea em todo o estado e alguns municípios possuem situações mais precárias, com índices de mortalidade mais altos.

Além disso, por mais baixo que sejam os índices, sempre é possível melhorá-los. Como afirma Ripsa (2002), "taxas reduzidas também podem encobrir más condições de vida em segmentos sociais específicos", o que indica oportunidades de atuação também em municípios com números considerados excelentes.

Outro ponto a ser considerado é que, enquanto se comemora apenas 1,2% de mortalidade infantil em determinada cidade, é esquecido o fato de que em um município como São Paulo, onde nasceram mais de 173 mil crianças em 2008, essa taxa reduzida significa a morte de mais de 2 mil crianças por ano. Quantas dessas mortes não poderiam ter sido evitadas?

Esses pontos já seriam suficientes para justificar uma abordagem inovadora dos dados do SINASC e SIM com o intuito de gerar informações que possam ser úteis no combate à mortalidade infantil.

Todavia, há um motivo adicional: quando as taxas de mortalidade infantil são altas, é fácil identificar ações para reduzi-las. Entretanto, em níveis baixos, estudos detalhados crescem em importância, uma vez que nesse cenário fica mais difícil propor ações assertivas.

2. OBJETIVOS

Os principais objetivos deste artigo podem ser assim sintetizados:

- Propor uma metodologia para cruzamento das bases de dados do Sistema de Informações sobre Nascidos Vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM), possibilitando a geração de uma base de dados consolidada que seja útil em estudos sobre mortalidade infantil.
- Exemplificar o potencial dessa metodologia utilizando a base consolidada para identificar os fatores individuais relacionados à mortalidade, no primeiro ano de vida, dos nascidos vivos no estado de São Paulo nos anos de 2006 e 2007.

O conhecimento gerado por essa abordagem permite a realização de diversos estudos, visando subsidiar processos de planejamento e gestão de políticas de saúde voltadas para a preservação da saúde infantil e redução da mortalidade nos primeiros anos de vida.

3. MORTALIDADE INFANTIL

A taxa de mortalidade infantil estima o risco de morte de crianças menores de um ano de idade e pode ser considerado um dos principais indicadores que refletem o nível de condições de vida em determinada unidade geográfica. Conceitualmente, a taxa de mortalidade infantil é definida como o "número de óbitos de menores de um ano de idade por mil nascidos vivos, na população residente em determinado espaço geográfico, no ano considerado" (RIPSA, 2002). Ela mede o risco que tem um nascido vivo de morrer antes de completar um ano de vida, fato que está ligado às condições de habitação, saneamento, nutrição, educação e também de assistência à saúde e principalmente ao pré-natal, ao parto e ao recém-nascido.

Ainda de acordo com Ripsa (2002), essa taxa pode ser utilizada para:

- Analisar variações geográficas e temporais da mortalidade infantil, identificando tendências e situações de desigualdade que possam demandar a realização de estudos especiais.
- Contribuir na avaliação dos níveis de saúde e de desenvolvimento socioeconômico da população.
- Subsidiar processos de planejamento, gestão e avaliação de políticas e ações de saúde voltadas para a atenção pré-natal e o parto, bem como para a proteção da saúde infantil.

Nas últimas décadas, a taxa de mortalidade infantil sofreu significativa redução no mundo graças aos meios mais eficazes de controle das doenças endêmicas, ao avanço da Medicina e às melhorias das condições gerais de bem-estar da população.

O Brasil também apresentou consideráveis avanços. Em 1991 foram registrados 45,1 óbitos de menores de um ano de idade por mil nascidos vivos. No ano 2000, o mesmo índice foi reduzido para 30,1 e, em 2009, a taxa foi de 22,5 óbitos para cada mil nascidos vivos (IBGE, 2011). Esses números refletem a melhoria nas condições de vida, o declínio da fecundidade e o efeito de intervenções públicas nas áreas de saúde, saneamento e educação, entre outros aspectos. Entretanto, as desigualdades no país ainda são evidentes e muitas regiões estão muito longe do patamar de países desenvolvidos, onde a taxa de mortalidade é inferior a 10 óbitos por mil nascidos vivos.

Segundo Waldvogel *et al*. (2008), o Brasil conta com duas fontes produtoras de dados sobre nascimentos e óbitos. A primeira, coordenada pelo IBGE, foi criada com base nas informações do registro civil de pessoas naturais, obtidas nos Cartórios de Registro Civil de todos os municípios brasileiros. A segunda, sob a coordenação do Ministério da Saúde e implantada em todas as Secretarias Municipais de Saúde do país, utiliza as informações contidas nas declarações de óbito (DO) e declarações de nascido vivo (DN).

Ainda de acordo com Waldvogel *et al*. (2008), o estado de São Paulo é a única unidade da federação que desenvolveu um sistema próprio de produção de estatísticas vitais que é independente e, ao mesmo tempo, integrado aos sistemas nacionais do IBGE e do Ministério da Saúde. Dessa forma, a Fundação Seade consegue tratar os dados epidemiológicos originários das DOs e DNs e também aqueles do registro civil e produzir bases mais abrangentes e consistentes.

As bases do SIM e SINASC das regiões Sudeste e Sul são consideradas mais precisas e menos sujeitas a erros (BRASIL, 2004). Assim, este estudo demonstra a utilização da metodologia de cruzamento dessas bases, limitando a análise apenas aos nascimentos e óbitos ocorridos no estado de São Paulo.

4. METODOLOGIA

4.1 Cruzamento dos bancos de dados: SINASC x SIM

Os dados do SINASC foram cruzados com os dados do SIM. O SIM contempla todas as declarações de óbito consolidadas na Coordenação-Geral de Informações e Análise Epidemiológica, do Departamento de Análise de Situação de Saúde, da Secretaria de Vigilância em Saúde. Todavia, neste estudo foram utilizadas apenas as declarações de óbitos de menores de um ano de idade.

O cruzamento dos dados teve como intuito a criação de uma base consolidada com informações dos dois sistemas. Ou seja, uma base contendo todos os dados a respeito do recém-nascido, da mãe, da gestação e do parto, bem como a identificação da sobrevivência ou falecimento de cada criança no seu primeiro ano de vida, e detalhes das respectivas causas de óbito quando este ocorreu. Essa base unificada permite estudar detalhadamente as variáveis que levam a fatores de risco para o recém-nascido.

Para marcar precisamente quais crianças evoluíram a óbito, é necessário ao menos um campo-chave que seja comum entre as duas bases. Felizmente, as DOs possuem um campo no qual deve ser preenchido o número da DN, que é o identificador das declaração de nascidos vivos. A partir desse campo foi possível cruzar os dados dos dois sistemas de modo a identificar os nascidos vivos que morreram no primeiro ano de vida.

Entretanto, nem todas as declarações de óbito de menores de um ano de idade estão com o número da declaração de nascido vivo (DN) preenchido. Nesses casos, não foi possível achar a correspondência exata desses óbitos na base de nascidos vivos. Este e outros problemas de qualidade dos dados poderiam gerar dois tipos de erros no processo de integração das bases. O primeiro diz respeito à marcação de falecimento equivocada, ou seja, indicar uma criança como morta, sendo que ela, na realidade, sobreviveu no seu primeiro ano de vida. O segundo caso é justamente o oposto, ou seja, marcar como sobrevivente uma criança que, na realidade, evoluiu a óbito. É possível fazer uma analogia desses possíveis erros de classificação com os apresentados por Hair *et al*. (2005a) ao discutir os tipos de erros de um processo de inferência estatística.

Dependendo do objetivo do estudo que fará uso dessa base consolidada, um tipo de erro pode ser mais tolerável que outro. A principal contribuição deste artigo é descrever um procedimento de unificação que prioriza eliminar o primeiro tipo de erro, bem como minimizar a possibilidade do segundo tipo, ainda que, para tanto, tal procedimento opte por eliminar os registros potenciais de erro do segundo tipo. O Quadro 1 sintetiza os tipos de erros que podem surgir na construção da base unificada e as abordagens adotadas em cada caso.

QUADRO 1. Possíveis erros no cruzamento das bases de dados.

Classificação do erro	Descrição	Abordagem pra minimizar o erro
Tipo I	Falso-positivo: indicação de óbito quando este não ocorreu.	Cruzamento das bases pelos campos: número da DN e data de nascimento.
Tipo II	Falso-negativo: indicação de não óbito quando este de fato ocorreu.	Exclusão de registros "semelhantes" aos óbitos não identificados.

Fonte: elaborado pelos autores.

Como indicado no Quadro, para evitar que entre os considerados "sobreviventes no primeiro ano de vida" houvesse casos de morte, optou-se por excluir todas as crianças com características semelhantes às de óbitos não identificados. Esse procedimento e outras consistências realizadas nas bases são descritos a seguir.

4.2. Preparação da base unificada

Segue-se uma descrição da preparação da base de nascimentos em 2006 com óbitos em 2006 ou 2007. O mesmo procedimento foi executado posteriormente para o ano de 2007.

A base de óbitos (SIM) indicava 8.298 óbitos em 2006 e 2007 de nascidos em 2006. Antes de cruzar os dados para identificar esses óbitos entre o total de 603.368 nascidos em 2006 (base de nascimentos do SINASC), foi necessário excluir as declarações de nascidos vivos que possuíam números de DN duplicados. Foram excluídas também as declarações de óbitos na mesma situação.

Após o cruzamento dos dados pelo número da DN, foi constatado que em alguns casos com correspondência exata desse número a data de nascimento divergia. Esses registros também foram excluídos, por serem potenciais casos do erro tipo I, ou seja, indicações de óbito quando este não ocorreu. Já os 3.813 nascimentos com correspondência exata na base de óbito por número da DN e data de nascimento foram marcados como casos que evoluíram a óbito. Os óbitos já identificados e todos os demais cuja data de nascimento não era no ano de 2006 foram excluídos da base de óbitos. A Tabela 1 resume a situação após essa primeira etapa do cruzamento:

TABELA 1. Valores após a 1ª etapa do cruzamento das bases (ano de referência: 2006)

Base de Nascimentos (SINASC)	
Total de nascimentos em 2006	601.178
Nascimentos que evoluíram a óbito	3.813
Nascimentos que não evoluíram a óbito	597.365
Base de óbitos (SIM)	
Óbitos ainda não identificados	4.398

Fonte: elaborado pelos autores.

Para identificar os demais 4.398 óbitos na base de nascimentos, foi necessário cruzar as duas bases por outras variáveis. Porém, as variáveis disponíveis não possibilitam a identificação exata e, na maioria dos casos, vários nascimentos aparecem relacionados a um único óbito. Como a correspondência é pouco precisa, marcar esses nascimentos como casos que evoluíram a óbito implicaria aumentar muito o erro do tipo I. Todavia, simplesmente ignorar esses óbitos não identificados deixando de marcá-los na base de nascimentos significaria assumir que se teria mais de 4 mil casos do erro tipo II. A solução escolhida foi não aumentar o erro tipo I, ou seja, não marcar esses casos na base de nascimentos e, ao mesmo tempo, tentar minimizar a possibilidade de erro do tipo II por meio da exclusão dos nascimentos que, de alguma forma, pudessem sugerir uma relação com algum dos óbitos ainda não identificados.

Visando minimizar o número de nascimentos excluídos para cada óbito ainda não identificado e também evitar distorções nas características da base final, foram aplicados inicialmente cruzamentos específicos, com muitas variáveis, tornando-os mais genéricos (com menos variáveis) à medida que restavam menos óbitos não identificados. O processo de cruzamento com variáveis diferentes e exclusão dos nascimentos e óbitos relacionados foi executado 12 vezes, até que não restassem registros na base de óbitos.

O resultado final desse processo é sintetizado na Tabela 2. Cada linha indica uma variável e cada coluna representa uma das 12 etapas do processo nos quais as variáveis indicadas com um "X" foram utilizadas no cruzamento. O número de nascimentos (DNs) e óbitos (DOs) eliminados em cada etapa aparece nas duas últimas linhas da tabela.

TABELA 2. Processo de exclusão de óbitos não identificados (ano de referência: 2006).

Variáveis	Etapas											
	1	2	3	4	5	6	7	8	9	10	11	12
dtnasc	X	X	X	X	X	X	X	X	X	X	X	X
codmunres	X	X	X	X	X	X	X	X	X	X	X	
codbaires	X	X	X	X	X	X	X	X	X			
peso	X	X	X	X	X	X	X	X		X		
sexo	X	X	X	X	X	X	X					
racacor	X	X	X	X	X	X						
codocupmae	X	X	X	X	X							
escmae	X	X	X	X								
gestacao	X	X	X									
parto	X	X										
gravidez	X											
DNs	29	0	2	15	28	235	241	3	5.798	937	41.922	83.396
DOs	29	0	2	15	28	235	241	3	1.161	910	1.712	62

Fonte: elaborado pelos autores.

Ao término do procedimento, foram obtidos os seguintes valores:

TABELA 3. Valores finais para o ano de 2006.

Base de Nascimentos (SINASC)	
Total de nascimentos em 2006	468.572
Nascimentos que evoluíram a óbito	3.813
Nascimentos que não evoluíram a óbito	464.759
Base de óbitos (SIM)	
Óbitos ainda não identificados	0

Fonte: elaborado pelos autores.

O mesmo procedimento para consistência e cruzamento das bases foi realizado para os nascimentos de 2007 e óbitos de 2007 e 2008. As duas bases resultantes foram, então, consolidadas. Alguns poucos registros ainda tiveram que ser excluídos, pois apresentaram mesmo número de DN em anos distintos. Os valores dessa base consolidada são apresentados na Tabela 4.

TABELA 4. Valores da base consolidada (nascimentos de 2006 e 2007).

Base Consolidada – Nascimentos de 2006 e 2007	
Total de nascimentos em 2006 e 2007	932.744
Nascimentos que evoluíram a óbito	6.886
Nascimentos que não evoluíram a óbito	925.858
% de crianças que evoluíram a óbito	0,74%

Fonte: elaborado pelos autores.

Quando considerados os dados originais dos dois anos, antes do procedimento de cruzamento descrito anteriormente, o percentual de crianças que tinham evoluído a óbito era de 1,36%. Após o processo de unificação, devido à impossibilidade de identificar com precisão todos os óbitos, esse mesmo percentual era de apenas 0,74% na base consolidada. Em outras palavras, o processo de exclusão de registros resultou em uma base com proporção de óbitos diferente da população original.

Esse desbalanceamento poderia impactar negativamente qualquer estudo que fizesse uso dessa base. Para tratar essa questão, foi extraída uma amostra aleatória simples (HAIR *et al*. 2005b) das crianças que não evoluíram a óbito, de modo a manter a mesma proporção da população original. Portanto, na base consolidada foram mantidos todos os registros com indicação de óbito e o percentual excedente de registros de sobreviventes foi excluído aleatoriamente. Dessa forma, a base final ficou com um número de nascimentos e óbitos em proporção equivalente à original, ou seja, algo em torno de 1,36%.

4.3. Exemplo de uso da base unificada para identificação de fatores de risco

Por meio do procedimento de unificação das bases descrito anteriormente foi possível a criação de um único banco de dados no qual as variáveis disponibilizadas correspondem às características da gestante, da gestação, do recém-nascido vivo e local do nascimento. A esses dados foi adicionado um "*flag*", que indica se o recém-nascido sobreviveu ou faleceu no período de um ano. Essas informações serviram para a construção de um modelo de predição a partir da metodologia de regressão logística, cujo resultado possibilitou verificar quais variáveis constituem-se em fatores de risco.

A regressão logística é um tipo de regressão formulada para prever e explicar o comportamento de uma variável categórica binária a partir de variáveis independentes que podem ser métricas ou categóricas. Apesar dessa técnica ser mais adequada para variáveis dependentes que possuam apenas dois grupos, é possível utilizá-la também em situações em que a variável dependente possui três ou mais grupos (HAIR *et al*., 2005a; CORRAR *et al*., 2007).

Além de permitir a classificação de fenômenos ou indivíduos em categorias específicas, a regressão logística permite, ainda, estimar a probabilidade de ocorrência de determinado evento ou de que um fenômeno venha a se enquadrar nessa ou naquela categoria. Ou seja, os resultados da variável

dependente permitem interpretações em termos de probabilidade e não apenas a classificação em categorias (CORRAR *et al*., 2007).

Essas características fazem da regressão logística uma técnica adequada ao objeto de estudo deste trabalho, uma vez que a variável dependente em análise é binária, ou seja, indica apenas se a criança evoluiu ou não a óbito. Existem outras técnicas com finalidades semelhantes, como, por exemplo, a análise discriminante. Todavia, uma das vantagens da regressão logística em relação à análise discriminante reside no fato de que a regressão logística é uma técnica muito mais flexível, com suposições mais brandas. Por exemplo, apesar de desejável, a normalidade dos dados de entrada (variáveis independentes) não é uma condição determinante para aplicação eficaz dessa técnica. A regressão logística não depende das mesmas suposições rígidas que a análise discriminante e é muito mais robusta quando tais pressupostos não são satisfeitos (HAIR *et al*., 2005a).

Os resultados do modelo logístico aplicado à base unificada são descritos a seguir. O *software* estatístico utilizado no desenvolvimento do modelo foi o SPSS versão 18.

5. ANÁLISE DOS RESULTADOS DO MODELO LOGÍSTICO

A variável dependente de um modelo multivariado é aquela que está sendo prevista ou explicada pelo conjunto de variáveis independentes (HAIR *et al*., 2005a). Na base de dados unificada a variável dependente é o *flag*, indicativo de óbito ou sobrevivência no primeiro ano de vida. Essa variável binária assume os seguintes valores: 0 = "não" e 1 = "sim". Dessa forma, "0" indica sobrevivência, enquanto "1" indica óbito.

As demais variáveis que constam na base unificada podem ser consideradas variáveis independentes, ou seja, aquelas selecionadas como previsoras e potenciais variáveis de explicação da variável dependente (HAIR *et al*., 2005a). As variáveis independentes utilizadas no modelo contemplam as seguintes dimensões:

- *Dados demográficos*: têm o objetivo de informar o local de nascimento do recém-nascido (hospital, domicílio, outro estabelecimento de saúde, etc.), bairro de nascimento, etc.
- *Dados da gestante*: têm o objetivo de traçar o perfil das mulheres a partir da idade, estado civil, escolaridade (anos de estudo), ocupação, quantidade de filhos (vivos/mortos) e o município de residência.
- *Dados da gestação*: informam o número de semanas de gestação, tipo de gravidez, tipo de parto e quantidade de consultas pré-natal realizadas.
- *Dados do recém-nascido*: permitem caracterizar os recém-nascidos com informações referentes à data de nascimento, horário, sexo, raça, existência de anomalia congênita e o tipo, peso e avaliação de saúde segundo a escala de Apgar (frequência cardíaca, respiração, tônus muscular rígido ou flácido, coloração da pele e irritabilidade reflexa – se a criança está ativa e reativa à manipulação).

Entre essas variáveis, algumas são de natureza qualitativa e outras quantitativas (contínuas). O primeiro passo para a elaboração do modelo logístico foi a análise exploratória univariada, sendo que, para as variáveis do primeiro grupo, foi analisada a frequência de cada categoria, enquanto que as variáveis do segundo grupo foram caracterizadas a partir de medidas de tendência central, desvio-padrão, mínimo e máximo e, ainda, as medidas de assimetria e achatamento para analisar a forma como as observações estão distribuídas.

O segundo passo foi a elaboração de uma análise bivariada entre a variável dependente e cada uma das variáveis independentes, com o intuito de buscar as primeiras evidências de quais variáveis estão de alguma forma relacionadas à alta probabilidade de óbito. Para compreender a relação existente entre a probabilidade de ocorrer um evento/resposta e a variabilidade das variáveis independentes, Hosmer e Lemeshow (2000) utilizam um processo de categorização dos dados. O objetivo da categorização é aperfeiçoar a interpretação dos parâmetros e obter mais assertividade na classificação obtida no modelo, principalmente quando o comportamento da variável independente se relaciona de forma não linear com o evento de interesse. As variáveis contínuas podem ser utilizadas como são, mas, ao serem categorizadas, é reduzida a influência de valores discrepantes. As variáveis discretas, ou já categorizadas, devem ser agrupadas em número de categorias suficiente para obter uma análise robusta. As variáveis que apresentam elevado número de agrupamento com algumas categorias que têm baixa frequência devem ser reagrupadas. Dessa forma, a análise bivariada foi útil também para a categorização adequada das variáveis independentes, sendo que o percentual de incidência de óbitos foi analisado para definir as faixas das categorias de cada variável.

Na regressão logística, quando há muitas variáveis independentes, pode ser difícil selecionar as mais adequadas ao modelo, devido à multicolinearidade. Existem vários métodos de seleção de variáveis que conduzem ao "melhor modelo" (MAROCO, 2010). As variáveis deste estudo foram selecionadas primeiramente pelo método de seleção *Forward*, baseado na razão de verossimilhança. Os parâmetros de teste de ajuste de cada etapa do *out-put* foram analisados e as variáveis com melhores ajustes foram incluídas no modelo final de regressão logística, processado com o método de seleção *Enter*.

A Tabela 5 mostra o resultado do modelo final de regressão logística, contendo: variáveis explicativas, as respectivas descrições, categorias de referência, os coeficientes do modelo, as estatísticas Wald, graus de liberdade, p-*values* e os intervalos de confiança.

TABELA 5. Resultado do Modelo de Regressão Logística.

Variáveis	Descrição das Variáveis	Categoria de Referência	β	Wald	df	Sig.	Exp(β)	95% C.I. for EXP(B)	
								Lower	Upper
SEXO(1)	Sexo do RN: Masculino	Feminino	0,189	32,962	1	0,000	1,208	1,133	1,289
ID_ANOMAL(1)	Há anomalias congênitas: Sim	Não	2,904	2.264,562	1	0,000	18,248	16,191	20,567
PESO_FAIXA	**Peso ao nascer**			**1.388,266**	**3**	**0,000**			
PESO_FAIXA(1)	< 2500 g	≥4000 g	1,996	203,053	1	0,000	7,358	5,592	9,682
PESO_FAIXA(2)	2500 - 2999 g		0,505	12,896	1	0,000	1,658	1,258	2,184
PESO_FAIXA(3)	3000 - 3999 g		0,183	1,753	1	0,186	1,201	0,916	1,576
IDADEMAE_FAIXA	**Faixa Etária da Mãe**			**16,296**	**5**	**0,006**			
IDADEMAE_FAIXA(1)	< 15 anos	> 44 anos	-0,696	6,397	1	0,011	0,499	0,291	0,855
IDADEMAE_FAIXA(2)	15 - 19 anos		-0,827	9,946	1	0,002	0,438	0,262	0,731
IDADEMAE_FAIXA(3)	20 - 34 anos		-0,805	9,604	1	0,002	0,447	0,269	0,744
IDADEMAE_FAIXA(4)	35 - 39 anos		-0,790	8,973	1	0,003	0,454	0,271	0,761
IDADEMAE_FAIXA(5)	40 - 44 anos		-0,598	4,786	1	0,029	0,550	0,322	0,940
APGAR_1MIN	**Índice de Apgar - 1º min.**			**1.465,312**	**3**	**0,000**			
APGAR_1MIN(1)	0 - 2	8 - 10	2,147	952,362	1	0,000	8,556	7,466	9,806
APGAR_1MIN(2)	3 - 4		1,855	975,277	1	0,000	6,395	5,692	7,184
APGAR_1MIN(3)	5 - 7		1,181	820,123	1	0,000	3,259	3,006	3,533
APGAR_5MIN	**Índice de Apgar - 5º min.**			**1.154,180**	**3**	**0,000**			
APGAR_5MIN(1)	0 - 2	8 - 10	2,776	737,644	1	0,000	16,048	13,135	19,608
APGAR_5MIN(2)	3 - 4		2,427	543,306	1	0,000	11,324	9,233	13,887
APGAR_5MIN(3)	5 - 7		1,171	561,586	1	0,000	3,226	2,928	3,554
CONSULTAS_QTDE(1)	Qtde. consultas de pré-natal: < 7	≥ 7	0,611	300,622	1	0,000	1,841	1,719	1,973
GESTACAO	**Período de Gestação**			**631,467**	**2**	**0,000**			
GESTACAO(1)	≤36 semanas	≥ 42 semanas	0,739	7,621	1	0,006	2,093	1,239	3,537
GESTACAO(2)	37 - 41 semanas		-0,414	2,423	1	0,120	0,661	0,393	1,113
ESCMAE(1)	Escolaridade da mãe: < 8 anos	≥8 anos	0,093	6,822	1	0,009	1,098	1,024	1,178
EST_CIV_MAE(1)	Estado civil da mãe: Solteira/Outros	Casada	0,102	8,026	1	0,005	1,107	1,032	1,188
Constant	Intercepto		-5,724	210,811	1	0,000	0,003		

a. Variable(s) entered on step 1: SEXO, ID_ANOMAL, PESO_FAIXA, IDADEMAE_FAIXA, APGAR_1MIN, APGAR_5MIN, CONSULTAS_QTDE, GESTACAO, ESCMAE, EST_CIV_MAE.

A análise da significância de cada coeficiente é feita com base no teste Wald, que testa a hipótese nula de que o coeficiente estimado é igual a zero. Os resultados indicam que, no nível de significância de 5%, os coeficientes das variáveis explicativas selecionadas são estatisticamente diferentes de zero. As variáveis do modelo e as respectivas categorias são altamente significantes (Sig. < 0,05), exceto as categorias "peso: 3000-3999 g" da variável peso ao nascer e "período: 37-41 semanas" da variável semanas de gestação. Todavia, o p-*value* global dessas duas variáveis também é significante.

Entre as variáveis que não foram inclusas no modelo final por apresentarem estatísticas Wald não significantes, destacam-se: raça/cor, local de nascimento, tipo de gravidez e tipo de parto.

Um modelo adequado é aquele cujo escore (probabilidade) calculado consegue distinguir os eventos, ou seja, os óbitos e sobreviventes. Existem vários métodos de avaliação de desempenho de um modelo, sendo que neste estudo foram adotados dois deles: o método da tabela de classificação e a curva ROC.

A matriz de classificação consiste em determinar um ponto de corte (*classification cutoff*) no escore final do modelo. Escores acima desse ponto de corte indicam a presença do evento/resposta de interesse (óbito) e os escores abaixo desse ponto indicam ausência (sobreviventes) (FÁVERO *et al*., 2009).

A matriz de classificação do modelo ajustado, utilizando um ponto de corte em 0,014, indica que o nível de acerto geral do modelo é de 91,3% (acurácia). Do total de nascidos vivos que não evoluíram a óbito, 91,4% foram classificados de maneira adequada e, do total de óbitos, 83,4% foram corretamente identificados.

O segundo método adotado para avaliar o desempenho do modelo é a curva ROC. A curva ROC é obtida por um gráfico que relaciona a especificidade (eixo x) e a sensibilidade (eixo y). A sensibilidade são os declarados como óbitos que foram classificados corretamente e a especificidade são os sobreviventes que foram classificados corretamente. Segundo Fávero *et al*. (2009), quanto maior a área abaixo da curva, maior é a capacidade do modelo em discriminar os eventos de interesse.

Em geral, para que um modelo seja considerado adequado, espera-se um valor acima de 0,7. Uma área acima de 0,9 é considerada excelente, como é o caso do modelo desenvolvido, que apresentou área sob a curva ROC de 0,925. O Gráfico 1 mostra que a curva está muito distante da diagonal, corroborando a conclusão de que o modelo possui excelente grau de discriminação do evento de interesse.

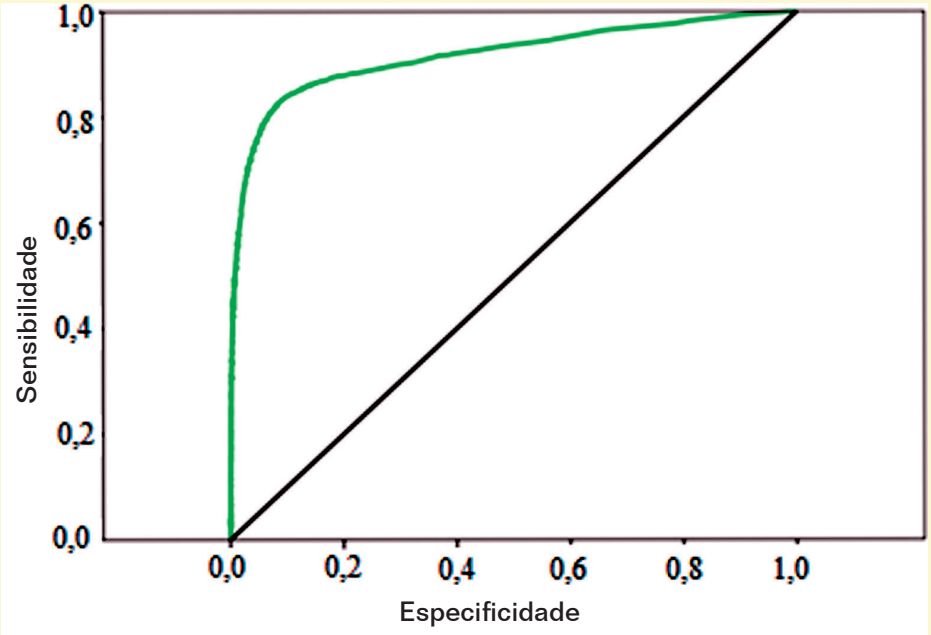

GRÁFICO 1. Curva ROC do modelo final.

Dado que o modelo mostrou-se adequado para prever a probabilidade de um recém-nascido evoluir a óbito no primeiro ano de vida, as variáveis a seguir, que compõem o modelo, podem ser caracterizadas como fatores de risco que levam a alto nível de mortalidade infantil.

Escolaridade da mãe
Quando o grau de escolaridade da mãe é abaixo de oito anos, o risco de morte do recém-nascido aumenta em relação àqueles cujas mães estudaram oito anos ou mais. As mães com alto grau de escolaridade podem pertencer às classes sociais mais favorecidas, deter mais conhecimentos sobre cuidados com o recém-nascido e direitos de cidadania e fazer prevalecer para si importantes exigências que podem impactar na saúde.

Estado civil
A categoria solteira/outros contribui para o aumento do risco de morte do recém-nascido em relação à categoria casada. O resultado evidencia que a ausência de um companheiro pode implicar dificuldades para dividir responsabilidades, custos financeiros e falta de apoio físico e moral.

Índice de Apgar no 1º minuto
É um fator de risco bastante expressivo. O risco de óbito é inversamente proporcional ao índice, quanto menor a faixa, mais alto o risco de óbito.

Índice de Apgar no 5º minuto
Semelhantemente ao índice anterior, também é fator de risco bastante expressivo. Quanto mais baixas as condições de vitalidade do recém-nascido nos primeiros minutos de vida, mais chances há de a criança evoluir a óbito.

Quantidade de consultas pré-natal
Há evidências de que quando a quantidade de consultas pré-natal é menor que sete, a chance do recém-nascido vivo evoluir a óbito aumenta consideravelmente em relação aos recém-nascidos de gestantes que fizeram mais de sete consultas. Segundo Ramos e Cuman (2009), a ausência de cuidados pré-natais está associada a risco de baixo peso ao nascer, partos prematuros e mortalidade materna e infantil.

Sexo
O coeficiente associado ao sexo indica que meninos possuem probabilidade um pouco mais alta de não sobreviverem ao primeiro ano de vida.

Período (semanas) de gestação
Naturalmente, a criança prematura apresenta mais riscos de morte no primeiro ano de vida. Quanto menor o número de semanas de gestação, maior esse risco.

Anomalia congênita
A presença de alguma anomalia congênita constitui importante causa de mortalidade infantil.

Peso ao nascer
A probabilidade de um recém-nascido com peso ao nascer inferior a 2.500 g morrer é consideravelmente maior quando comparada à probabilidade dos recém-nascidos com peso igual ou acima de 4.000 g (categoria de referência). Nota-se significativo aumento na probabilidade de ocorrer óbito à medida que o peso ao nascer diminui. O baixo peso ao nascer é impacto da desnutrição e reflete as condições nutricionais tanto do recém-nascido como da gestante (RAMOS; CUMAN, 2009).

Idade da gestante
O risco de evolução a óbito é mais alto quando as mães são muito jovens ou quando possuem idade avançada. As razões do elevado risco em mães jovens podem ser oriundas de implicações biológicas, familiares, emocionais ou econômicas. As mães com idade avançada têm mais chances de apresentarem problemas de saúde como diabetes, hipertensão arterial, placenta prévia e outros fatores que podem estar relacionados ao óbito infantil.

6. CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma proposta metodológica de integração das bases de dados do SINASC e SIM de modo a criar uma terceira base consolidada com informações de ambos os sistemas. Essa base integrada contém uma amostra representativa dos dados originais e pode ser útil em estudos sobre mortalidade infantil.

Para ilustrar a utilidade dessa metodologia, foi elaborado um modelo de regressão logística com o objetivo de identificar, entre as variáveis constantes nas declarações de nascidos vivos do SINASC, quais representam fatores de risco que podem levar um recém-nascido a falecer ainda no primeiro ano de vida.

A anomalia congênita, peso ao nascer e índice de Apgar no 1º e 5º minuto são os fatores que apresentaram mais impacto na mortalidade infantil. Entretanto, merece atenção especial o elevado número de óbitos dos recém-nascidos com peso abaixo de 2.500 g. Esse fator pode ser reflexo das condições nutricionais tanto do recém-nascido como da gestante e influencia o crescimento e desenvolvimento da criança. Trabalhos como o da Pastoral da Criança são fundamentais para sanar esse tipo de problema.

Considerando o perfil da mortalidade infantil detectado neste estudo, a acessibilidade e aprimoramento da assistência pré-natal e a assistência aos recém-nascidos reduzem a proporção de nascidos vivos com baixo peso, facilitam a detecção das gestações com desenvolvimento insatisfatório (anomalia congênita) e reduzem o número de óbitos decorrentes de afecções do período perinatal.

Entre as limitações deste estudo, pode-se citar o fato de que a metodologia proposta para cruzamento das bases pode, eventualmente, gerar uma amostra com algum viés em relação à população original. Isso pode ocorrer, por exemplo, caso haja uma padrão entre as declarações de óbito com e sem número de DN preenchido ou, ainda, caso o processo de exclusão das DNs elimine proporção muito grande de um grupo de crianças com determinada característica. Todavia, a interpretação do resultado do modelo logístico é consistente com o que era esperado, de modo que é pouco provável que tenha ocorrido essa "contaminação" na base unificada.

Outros estudos poderiam ser realizados com essa mesma base. Tais estudos poderiam considerar a divisão entre mortes neonatal (0 a 27 dias) e pós-neonatal (28 dias a menos de um ano), uma vez que a maior parte das mortes ocorre no período neonatal. Trabalhos futuros poderiam analisar também as causas de morte que constam no sistema do SIM, identificando fatores de risco para causas específicas.

Considerando o grande volume de dados secundários disponíveis, não apenas na área de saúde, mas também em outras áreas como educação e segurança, por exemplo, este trabalho pode inspirar outros pesquisadores a olharem para esses dados com outros olhos, procurando formas inovadoras de cruzar informações, identificar padrões, etc., sempre com o intuito de gerar conhecimento útil para a tomada de decisão.

REFERÊNCIAS

BRASIL. Ministério da Saúde. *Sistemas de Informações sobre Mortalidade (SIM) e Nascidos Vivos (SINASC) para os profissionais do Programa Saúde da Família*. 2 ed. Brasília: Ministério da Saúde, 2004.

CORRAR, L. J. *et al*. (Org.). *Análise multivariada*: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007.

DATASUS. Departamento de Informática do SUS – Datasus. *Sistema de Informações sobre Nascidos Vivos – SINASC*. Disponível em: <http://tabnet.datasus.gov.br/tabdata/sinasc/ dados/nov_indice.htm>.

DATASUS. Departamento de Informática do SUS – Datasus. *Sistema de Informações sobre Mortalidade – SIM*. Disponível em: <http://tabnet.datasus.gov.br/tabdata/sim/dados/cid10_indice.htm>.

FÁVERO, L. P.; BELFIORE, P. *Análise de dados*: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

SEADE. Fundação Sistema Estadual de Análise de Dados. *Informações dos Municípios Paulistas – IMP*. Disponível em: <http://www.seade.gov.br/produtos/imp/>.

HAIR JR, J. F. *et al*. *Análise multivariada de dados*. 5 ed. Porto Alegre: Bookman, 2005a.

HAIR JR, J. F. *et al*. *Fundamentos de Métodos de Pesquisa em Administração*. Porto Alegre: Bookman, 2005b.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. 2 ed. New York. Wiley, 2000.

IBGE. *Estudos e Análise da Dinâmica Demográfica – GEADD*. Censos Demográficos e pesquisas por amostragem. Disponível em: <http://seriesestatisticas.ibge.gov.br/series.aspx? vcodigo=IU32&sv=94&t=taxa-de-mortalidade-infantil/>.

MAROCO, J. *Análise Estatística com utilização do SPSS*. 3 ed. Lisboa: Silabo, 2010.

MORAIS NETO, O. L. *A mortalidade infantil no município de Goiânia*: uso vinculado do SIM e SINASC. Campinas, 1996. Dissertação (Mestrado em Saúde Coletiva) – Faculdade de Ciências Médicas, Unicamp, 1996.

RAMOS, H. A. C.; CUMAN, R. K. N. *Fatores de Risco para prematuridade*: pesquisa documental. 2009. Disponível em: <http://www.scielo.br/pdf/ean/v13n2/v13n2a09.pdf>.

RIPSA. Rede Interagencial de Informações para a Saúde. *Indicadores básicos de saúde no Brasil*: conceitos e aplicações. Rede Interagencial de Informações para a Saúde – Ripsa. Brasília: Organização Pan-Americana da Saúde, 2002. 299 p.

WALDVOGEL, B. C. *et al*. *Base Unificada de Nascimentos e Óbitos no Estado de São Paulo*: instrumento para aprimorar os indicadores de saúde. In: XVI Encontro Nacional de Estudos Populacionais, ABEP. Caxambú, 2008.