InfoSAS: um sistema de mineração de dados para controle da produção do SUS

Osvaldo Carvalho
é doutor de Estado em Ciência da Computação pela Universidade Pierre et Marie Curie e professor associado da UFMG. Trabalhou com algoritmos distribuídos, e hoje leciona Programação de Computadores e desenvolve sistemas de informação.

Marcos Prates
é bacharel pela UFMG (2006) em Matemática Computacional, mestre pela UFMG (2008) em Estatística e PhD em Estatística pela University of Connecticut, EUA (2011). Desenvolve métodos estatísticos e algoritmos para análise de estatística espacial e aprendizado de máquina.

Raquel Minardi
é bacharel em Ciência da Computação, doutora pela UFMG e pós-doutora pelo Comissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA). Desenvolve modelos e algoritmos em visualização de dados e em biologia computacional. Publicou mais de 20 artigos em periódicos internacionais e em conferências nacionais e internacionais.

Wagner Meira Jr.
é PhD pela Universidade de Rochester e professor titular de Ciência da Computação na Universidade Federal de Minas Gerais. Suas áreas de interesse são mineração de dados, sistemas paralelos e distribuídos e suas aplicações.

Renato M. Assunção
é PhD em Estatística pela University of Washington/ EUA e professor titular do Departamento de Ciência da Computação da UFMG, além de especialista em desenvolvimento de algoritmos e métodos para a análise de dados estatísticos, especialmente aqueles com georreferenciamento.

José Nagib Cotrim Árabe
é PhD em Ciência da Computação pela University of California, Los Angeles (UCLA), além de professor da Universidade Federal de Minas Gerais e chefe do Departamento de Ciência da Computação.

RESUMO

Este trabalho apresenta o InfoSAS, um sistema de detecção de anomalias estatísticas nos registros da produção do SUS. O InfoSAS encontra taxas de atendimentos por habitante muito superiores à média nacional, ou valores de internação bem acima dos praticados pela maioria dos estabelecimentos para um mesmo procedimento. Os resultados encontrados indicam que centenas de milhões de reais gastos pelo SUS são destinados a atendimentos considerados anômalos por critérios conservadores. Anomalias estatísticas podem ser provocadas por fraudes, mas também por mutirões de saúde, epidemias, ou má distribuição do atendimento. Em qualquer caso, anomalias graves devem ser investigadas ou explicadas.

Palavras-chave: SUS, Sistema Único de Saúde, Detecção de anomalias, Mineração de dados, Taxa de atendimentos por habitante, Valor médio de internações, Distribuição log-normal, Priorização de auditorias.

1. INTRODUÇÃO

A grande dimensão do setor de saúde e o enorme volume de recursos envolvidos fazem dele um alvo atraente para fraudes em todo o mundo. Nos Estados Unidos, por exemplo, estima-se que mais de 270 bilhões de dólares sejam perdidos anualmente com fraudes (THE ECONOMIST, 2014). Não temos motivos para acreditar que no Brasil a situação seja diferente. Neste documento apresentamos o InfoSAS, um sistema de mineração e visualização de dados do SUS que, como diversos outros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), identifica padrões e modelos variados nos dados, podendo ser de grande valia para a gestão do SUS e, em particular, para a identificação de anomalias estatísticas que podem ser indícios de irregularidades. O foco deste artigo é a sua utilização para fins de auditoria.

O InfoSAS encontra em profusão fatos que merecem a atenção de gestores, como taxas de atendimentos por habitante muito superiores à média nacional, como se pode ver na Figura 2; ou valores de internação bem acima dos praticados pela maioria dos estabelecimentos, como mostrado na Figura 4. Temos elementos, expostos na Seção 5, para acreditar que sua utilização na priorização de auditorias pode gerar retornos para o SUS da ordem de R$400 milhões por ano. Não é exagero estimar em bilhões de reais o valor que pode ser recuperado com auditorias a serem feitas sobre a produção de um período de 8 anos (de 2009 a 2016), assim como com a inibição de comportamentos anômalos na produção futura.

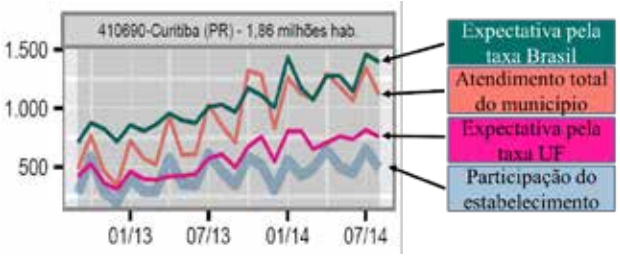
2. DISCREPÂNCIAS ESTATÍSTICAS

Os atendimentos prestados pelo SUS (denominados conjuntamente de produção do SUS) são registrados por diversos instrumentos e esses registros são armazenados nas bases SIA[1] e SIH[2]. O SUS também mantém

a base CNES[3], um cadastro de estabelecimentos. O sistema InfoSAS examina essas bases do SUS e também dados populacionais fornecidos pelo IBGE, e produz o que chamamos de folhas de fatos.

As bases de dados referidas são volumosas e essas folhas de fatos dão foco ao seu exame. Elas relacionam um alvo de mineração, um período de análise e um estabelecimento prestador, mostrando gráficos e tabelas obtidas pelo exame das bases de dados. Um alvo de mineração é um subconjunto da produção do SUS, definido por um ou mais procedimentos da Tabela do SUS[4], e algumas vezes também pela faixa etária dos pacientes. É importante observar que fatos presentes nas folhas podem também ser verificados de forma independente consultando outras fontes de informação oferecidas pelo SUS, como o TabWin[5].

Figura 1:
Análise da participação de um estabelecimento no atendimento de Curitiba para o alvo "Tratamento de doenças do aparelho da visão"



A Figura 1 mostra um exemplo de gráfico presente em uma folha de fatos. Nele vemos que, para o alvo "Tratamento de doenças do aparelho da visão", o atendimento total de Curitiba, no período de setembro de 2013 a agosto de 2014, seguiu de perto a expectativa segundo a média brasileira, que este atendimento é superior à média do Paraná, e que o prestador contribuiu com menos da metade desse atendimento. Essas são informações compatíveis com uma situação de normalidade.

Mas nem sempre é assim. Alguns fatos chamam atenção do profissional de controle. Na Figura 2, diversos municípios estão com uma taxa de atendimentos por habitante muito acima da brasileira e da estadual, e o estabelecimento em questão é praticamente o único a atender esses municípios. Este é um exemplo claro do que consideramos anomalias ou discrepâncias estatísticas.

Discrepâncias estatísticas devem ser consideradas com cuidado, como ilustra o exemplo mostrado na Figura 3. Cada um dos seis gráficos mostrados se refere a um município e todos a um único prestador. Pode-se notar que somente o gráfico no canto superior esquerdo (que corresponde ao município sede do estabelecimento) não apresenta picos estranhos, com valores muito acima dos esperados pelas taxas brasileiras e da UF onde se localizam os municípios. Um indício de fraude? Não. Quando se nota que o alvo é "Mamografia" e que o prestador é uma unidade móvel, a anomalia estatística se explica: o prestador é um veículo equipado com um

Figura 2:
Análise da taxa de atendimento por habitante de 6 municípios atendidos por um estabelecimento para o alvo "Tratamento de doenças do aparelho da visão"

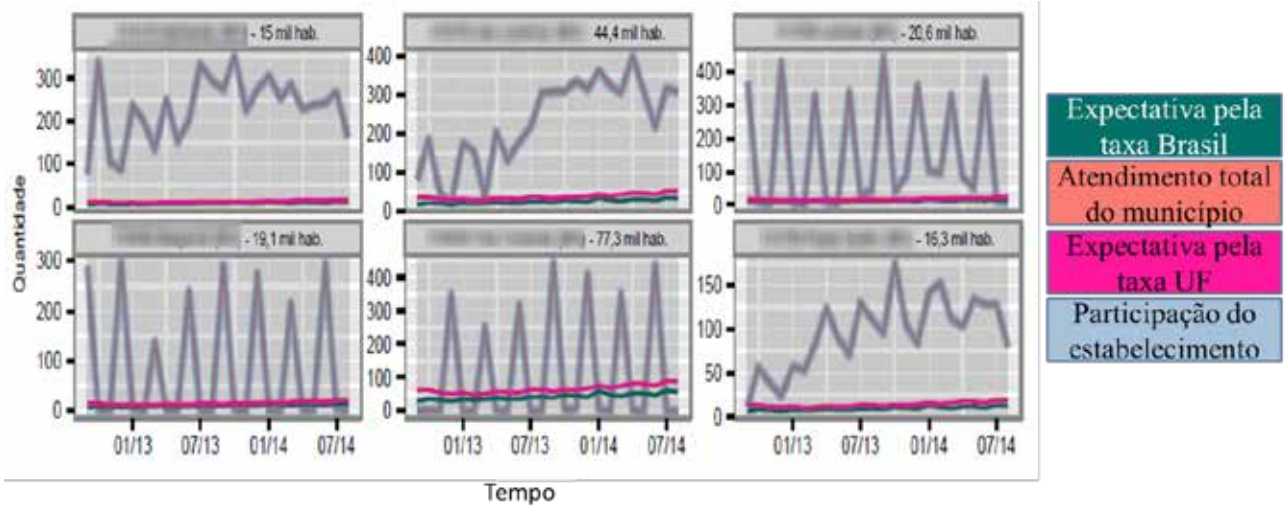
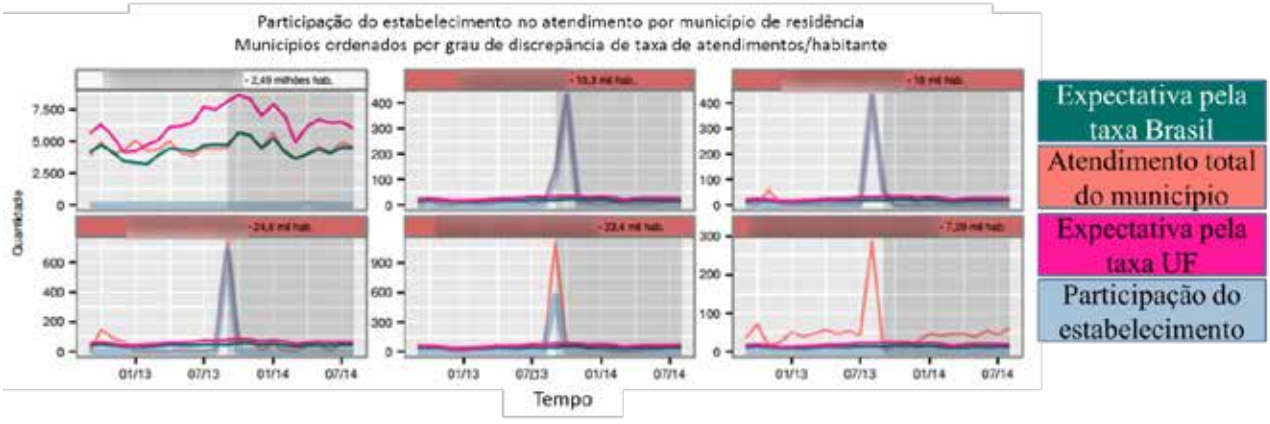


Figura 3:
Discrepâncias estatísticas explicáveis pelas características do alvo, "Mamografia bilateral para rastreamento", e do prestador, uma Unidade Móvel de Saúde da Mulher

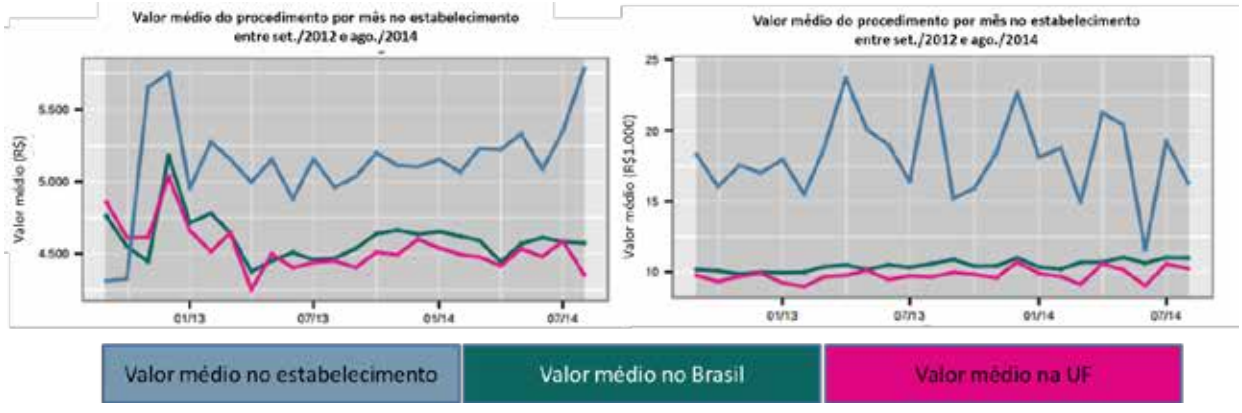


mamógrafo que realiza mutirões em pequenas cidades, provocando em suas visitas os picos de atendimentos.

De forma geral, é importante observar que uma anomalia estatística pode sim ser um indício de fraude, mas também pode resultar de processos corretos, como no exemplo acima, ou de informação incorreta, ou de epidemias. E é também possível que municípios com taxas de atendimento altas sejam, na verdade, os poucos bem atendidos no país. Portanto, a análise de uma folha de fatos deve ser feita com o máximo de discernimento e por um profissional de saúde pública que conheça o contexto local.

Na Figura 4 temos outros exemplos de discrepâncias estatísticas, agora com estabelecimentos praticando preços médios de internação muito superiores aos praticados no restante do Brasil para dois alvos de mineração. Novamente, é preciso que o gestor examine com cuidado cada caso de discrepância estatística, pois é possível que o estabelecimento tenha um perfil de pacientes com complicações acima da média, ou que utilize práticas que podem elevar o custo da internação, mas que também, digamos, diminuam sua taxa de óbitos.

Figura 4:
Análise do valor cobrado por um estabelecimento para internações dos alvos "Artroplastia total primária do quadril não cimentada/híbrida", à esquerda, e por outro estabelecimento para "Cirurgia cardiovascular – Cirurgia cardiovascular marcapasso", à direita



3. DETECÇÃO AUTOMÁTICA DE DISCREPÂNCIAS ESTATÍSTICAS

Encontrar folhas de fatos de interesse para o controle em uma Secretaria de Saúde é um fator de convencimento para a decisão de apuração. Entretanto, buscar ao acaso discrepâncias relevantes é como procurar uma agulha em um palheiro. Temos mais de 5000 alvos, aproximadamente 6000 prestadores e, considerando apenas períodos de 12 meses, em 3 anos de produção, seriam 36 janelas de tempo possíveis. Um cálculo simples mostra que temos, literalmente, bilhões de folhas de fatos que podem ser extraídas das bases de dados examinadas.

O InfoSAS mostra seu valor nesse momento, pois utiliza diversos algoritmos que procuram capturar discrepâncias, produzindo escores que permitem ordenação e priorização das folhas de fatos. Para a seleção, o InfoSAS também permite ao usuário concentrar-se em áreas definidas por filtros geográficos, por período de análise e por alvos, pois um profissional de controle e avaliação tem, muitas vezes, sua atenção dirigida para setores específicos da saúde, como cardiologia ou ortopedia e, claro, maior interesse em sua região de atuação.

O InfoSAS analisa as séries temporais de valor médio mensal por procedimento e de produção mensal em cada alvo desejado. Tais séries são calculadas por estabelecimento e por município de residência dos pacientes. Vários algoritmos de detecção de anomalias são utilizados e cada um deles calcula um escore. Estes escores são posteriormente combinados. Por falta de espaço para descrever em detalhes todos os escores, vamos apresentar a seguir a definição de dois deles. Uma descrição mais detalhada de outro algoritmo pode ser encontrada em (CARVALHO et al., 2015).

Um algoritmo tem como objetivo detectar oscilações abruptas na produção de um prestador. Para isso, a produção do estabelecimento é comparada com sua própria série histórica de produção (ou de valor médio). Se houver uma mudança brusca na série, o algoritmo atribui um escore de discrepância ao estabelecimento. Mais especificamente, em cada mês i e para cada estabelecimento l, nós calculamos $escore_{li}=(t_{li}+1)/(mediana_{li}+1)$, onde t_{li} é a dimensão de interesse (que pode ser o valor ou o número de atendimentos) para o l-ésimo estabelecimento no i-ésimo mês, $mediana_{li}$ é a mediana dos últimos m meses para a quantidade de interesse do l-ésimo estabelecimento no i-ésimo mês. Tipicamente, tomamos m=12 ou m=6 meses.

Outro algoritmo procura detectar discrepâncias em taxas de atendimento por habitante. Tomando como base a produção brasileira por 100 mil habitantes, a série

temporal de produção realizada por todos os estabelecimentos nos residentes de um município é analisada. Caso este município tenha uma produção acima do limiar, o algoritmo atribui um escore de discrepância a ele. Em seguida, o escore é atribuído aos estabelecimentos de forma proporcional à participação de cada um no atendimento a esse município.

O escore do estabelecimento é obtido somando-se os escores que ele obteve em cada município de atuação. Mais formalmente, o escore de produção do município é calculado como uma soma acumulada nos últimos 12 meses, $escore_{li}=\sum_{j=i-11}^{i} dif_{lj}$, onde $dif_{li}=max\{0,tBayes_{li}-limiar_i\}$ e $limiar_i=k*taxaBrasil_i$. O valor de $taxaBrasil_i$ é a taxa de produção mensal brasileira por 100 mil habitantes no mês i, $tBayes_{li}$ é a taxa bayesiana empírica por 100 mil habitantes no mês i para o município l, e k é uma constante previamente definida. Em nossos estudos, temos usado k=3. A taxa bayesiana empírica (ASSUNÇÃO et al., 1998; MARSHALL, 1991) é uma técnica estatística para calcular taxas e razões que não é afetada pela flutuação de pequenas populações.

4. FUNCIONAMENTO DO INFOSAS

A Figura 5 mostra o fluxo de dados do InfoSAS. Mensalmente, dados das bases SIA, SIH e CNES, e tam-

bém dados populacionais do IBGE alimentam o servidor de mineração de dados, que é guiado por uma tabela de alvos de mineração e que executa algoritmos para detecção de discrepâncias estatísticas. Esta fase de mineração produz o que nós chamamos de cubo de fatos minerados. Este cubo é explorado pelo usuário do InfoSAS, utilizando uma ferramenta de BI. O usuário seleciona um modelo de relatório e especifica parâmetros para filtros por conjunto de alvos, período de análise e recortes geográficos. Um relatório é produzido de acordo com o modelo escolhido e com os filtros determinados, contendo escores de discrepância calculados pelos algoritmos de mineração. Do relatório, o usuário extrai para análises mais aprofundadas folhas de fatos que lhe chamam a atenção.

5. CONCLUSÃO E TRABALHOS FUTUROS

No momento atual (outubro de 2016), o InfoSAS está instalado e em funcionamento no DATASUS. Um curso a distância para capacitação de gestores quanto ao uso do InfoSAS está em preparação, devendo ser finalizado até novembro de 2016, com a primeira oferta prevista ainda para este ano. O sistema InfoSAS já foi apresentado em diversos eventos nacionais (DRAC SAS, 2015) e, em todos eles, seus resultados foram julgados muito interessantes por especialistas em saúde pública.

Os resultados já obtidos pelo projeto InfoSAS são importantes e representam um passo à frente na modernização dos processos de seleção de itens para auditoria e controle. Além da necessidade de manutenção do sistema, constante atualização, correção e calibragem fina dos algoritmos já utilizados, existem muitas linhas para desenvolvimentos futuros. Vislumbramos uma nova interface com visualização mais flexível e interativa dos resultados; uma caracterização estatística de alvos com identificação de vazios assistenciais; a utilização da população SUS nos cálculos estatísticos; o cálculo de limites inferiores e superiores para a atribuição de produção anômala de um município a seus prestadores; a escolha adaptativa de recortes geográficos com o dimensionamento adequado à frequência de ocorrência de cada alvo; e a utilização de paralelismo no processamento da mineração e da visualização.

Além desses, queremos destacar dois desenvolvimentos que, a nosso ver, são os mais importantes a serem feitos. O primeiro destes seria viabilizado pela realização de auditorias, talvez guiadas por alertas do InfoSAS, classificando atendimentos um a um como fraudulentos ou regulares. Essa rotulação abre a possibilidade de utilização de algoritmos de aprendizado supervisionado, permitindo a emissão de alertas de forma mais precisa e elaborada.

O outro consiste em uma análise consolidada de anomalias em taxas de atendimentos por habitante. O InfoSAS, em sua versão atual utiliza diversos algoritmos que procuram capturar anomalias estatísticas. Temos

Figura 5:
Etapas da análise de dados pelo sistema InfoSAS

também uma estrutura de alvos de mineração onde um alvo pode incluir outros alvos. Com isso, conseguimos muitos achados, mas não temos como classificar estabelecimentos pelo conjunto de alvos nem obter estimativas globais de valor para produção anômala.

Para produzir um novo relatório superando estes problemas, nós decidimos ter como alvos somente conjuntos de procedimentos em uma mesma *forma de organização* da Tabela do SUS, o que faz com que a interseção entre dois alvos quaisquer seja vazia, e utilizar um único algoritmo com uma metodologia estatística para estimativa de excessos em taxas de atendimentos por habitante. Isso torna possível a obtenção de estimativas globais (em todos os alvos) de produção anômala para cada prestador, o que permite a priorização da apuração de fatos pelos gestores do SUS.

A metodologia que pretendemos utilizar, e que já aplicamos a dados de 2013, tem como ponto de partida a constatação de que a distribuição das taxas de atendimento por habitante observadas nos municípios segue uma distribuição log-normal, como ilustrado na Figura 6. Isso abre a possibilidade de definição para qualquer alvo e para qualquer município de um ponto de corte para o que deve ser considerado como normal ou anormal em uma taxa de atendimentos por habitante.

Nós arbitramos que taxas de atendimento por habitante devem ser consideradas anormais quando se situarem à direita do ponto que divide a área sob a curva em uma parte "normal", com 99% de probabilidade; e em uma parte "anormal", com 1% de probabilidade, o que nos parece um critério bastante prudente. Para o exemplo da Figura 7, taxas acima de 3,2 atendimentos por 1.000 habitantes/mês são consideradas anormais. Suponhamos que para o alvo em questão, um município de 100.000 habitantes tenha recebido 400 atendimentos em um dado mês. Pela taxa limite de 3,2 atendimentos/1.000 habitantes, a população desse município deveria ter recebido, no máximo, 320 atendimentos. Nós consideramos, então, que o município teve 80 atendimentos anômalos, cujo valor é distribuído entre os prestadores do município, mantendo as proporções do atendimento de cada prestador.

Os resultados da aplicação desta análise à produção registrada no SUS em 2013 são fortes. De um total de R$19.912.491.904,00, foram considerados anormais atendimentos com valor somado de R$413.920.365,56, correspondendo a 2,08% do total. O excesso em alvos registrados no SIHSUS foi de R$350.354.969,25, e no SIA de R$63.565.396,30.

Outro resultado deste estudo que pode ser utilizado para priorização de auditorias é uma análise da distribuição entre prestadores dos atendimentos considerados anômalos. Em todo o Brasil, apenas 5 prestadores concentram quase 9% do valor total das anomalias;

Figura 6:
Densidade da probabilidade de taxas de atendimento por habitante: log-normal teórica vs. dados observados

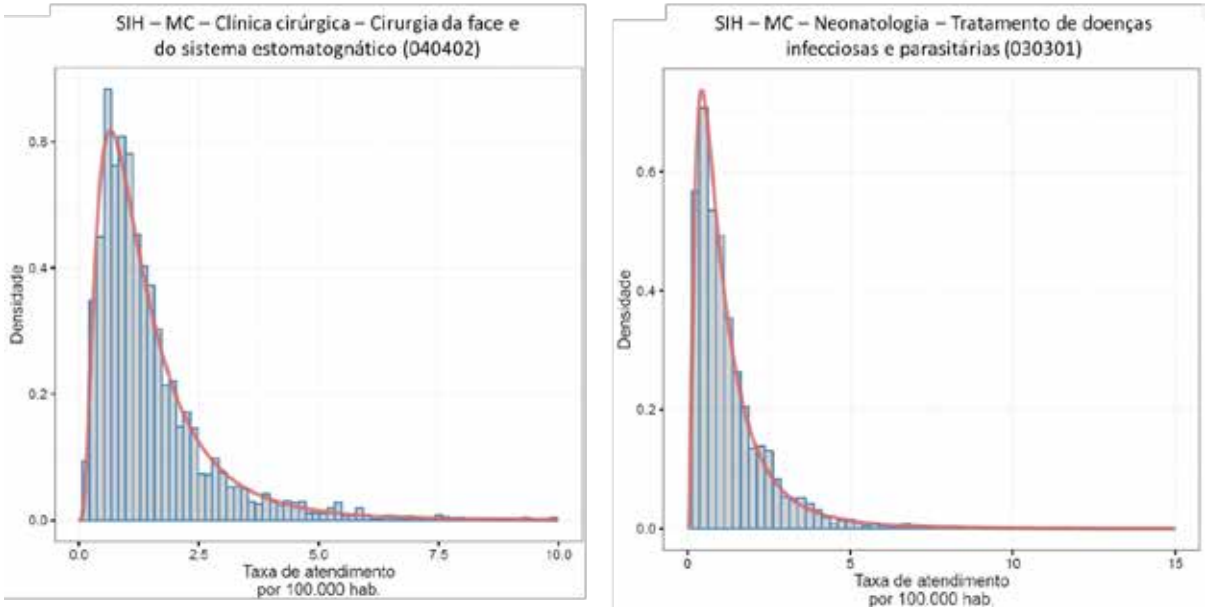
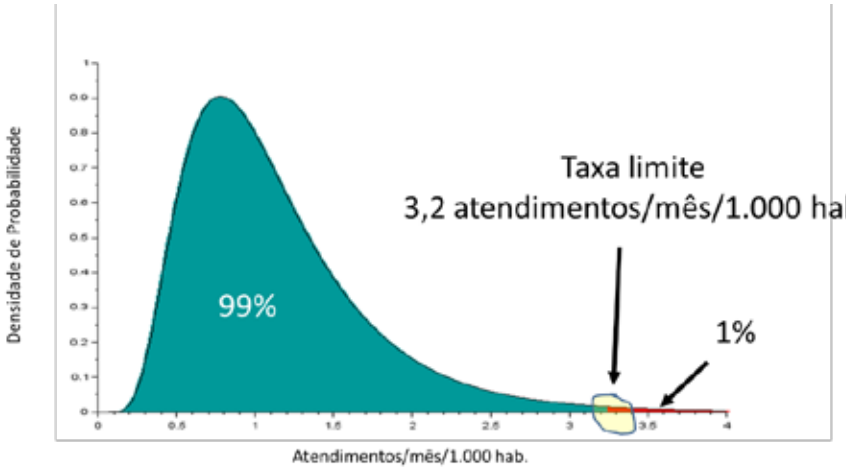


Figura 7:
Definição da taxa-limite de atendimentos por habitante no ponto de corte de 1% da distribuição log-normal



mais de 50% desse total está concentrado em 100 prestadores. Um único prestador teve quase 10 milhões de reais em atendimentos considerados como anômalos em 2013.

6. AGRADECIMENTOS

Em primeiro lugar temos que agradecer ao Departamento de Regulação, Avaliação e Controle da Secretaria de Atenção à Saúde do Ministério da Saúde, que solicitou e financiou todo o projeto.

O InfoSAS foi construído por uma equipe grande e multidisciplinar. Maria Helena Brandão foi a condutora do projeto por parte do DRAC. Ester Dias, Marcelo Campos, Sônia Gesteira e Suzana Rattes, da SMSA-BH, Luciana Morais, da Funed, e Mônica Castro, da Unimed-BH, foram nossos consultores sanitaristas. Fabiana Peixoto e Letícia Neto foram as gerentes, e Tomas Schweizer foi o líder técnico. Edré Moreira, Carlos Teixeira, Douglas Azevedo, Larissa Santos, Luiz Fernando Carvalho, Luiz Gustavo Silva, Maurício Nascimento Jr., Milton Ferreira, José Carlos Serufo Jr., Pablo Fonseca, Renan Xavier e Raquel Ferreira, bolsistas de pós-graduação, participaram decisivamente na pesquisa, construção e implementação de diversos algoritmos. Felipe Caetano, Ícaro Braga, Geraldo Franciscani, João Paulo Pesce, João Victor Bárbara, Raphael de Faria e Wicriton Silva cuidaram de desenvolvimento, testes, visualização e bancos de dados. A toda essa equipe, verdadeira responsável pela construção do sistema, vão os nossos mais sinceros agradecimentos.

REFERÊNCIAS

ASSUNÇÃO, R. M. et al. Maps of epidemiological rates: a Bayesian approach. Cadernos de Saúde Pública, Rio de Janeiro, v. 14, n. 4, p. 713–723, out./dez. 1998.

CARVALHO, L. F. M. et al. A simple and effective method for anomaly detection in healthcare. In: 4TH WORKSHOP ON DATA MINING FOR MEDICINE AND HEALTHCARE, IN CONJUNCTION WITH THE 15TH SIAM INTERNATIONAL CONFERENCE ON DATA MINING, 2015, Vancouver, maio 2015. Disponível em http://homepages.dcc.ufmg.br/~carlos/papers/sdm/dmmh2015.pdf. Acesso em 29 nov. 2016.

DRAC SAS. Ministério da Saúde. Ciclo de Oficinas do DRAC – Controle de Avaliação. Brasília, DF, 1 set. 2015. Disponível em: <https://www.youtube.com/watch?v=vIaR_Q7T-Us>. Acesso em: 4 jul. 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI magazine, California, v. 17, n. 3, p. 37-54, 1996.

MARSHALL, R. J. Mapping disease and mortality rates using empirical Bayes estimators. Applied Statistics, Auckland, v. 40, n. 2, p. 283–294, 1991.

THE ECONOMIST. The $272 billion swindle. Londres, 31 maio 2014. Disponível em http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle. Acesso em: 29 nov. 2016.