SISTEMA DE INFORMAÇÃO EM SAÚDE SILVESTRE – "SISS-GEO"

Marcia Chame1, Helio J. C. Barbosa2, Luiz Gadelha2
Douglas A. Augusto1, Eduardo Krempser2, Livia Abdalla1

1. SOLUÇÕES COMPUTACIONAIS PARA OS DESAFIOS DA MODELAGEM DE EMERGÊNCIA DE DOENÇAS ORIUNDAS DA FAUNA SILVESTRE

As alterações ambientais, incluindo as mudanças climáticas e a perda da biodiversidade, são fatores determinantes para a emergência de doenças oriundas de animais silvestres [6] e podem estar na origem das forças seletivas de novas variações genéticas que permitem o rompimento de barreiras biológicas por agentes patogênicos e o aumento do potencial de dispersão de doenças em humanos. Embora não consideradas adequadamente nas políticas de vigilância em saúde, o quadro é relevante, uma vez que a maioria (60,3%) das doenças infecciosas circula entre animais e humanos (zoonoses), das quais 71,8% dessas são causadas por patógenos com origem na vida silvestre [12].

Essas emergências quase sempre estão associadas aos territórios mais atingidos por impactos naturais e antropogênicos, compondo também a gama de parâmetros que tornam as desigualdades sociais ainda mais severas e injustas, como forte repercussão e custos para a saúde e a qualidade de vida (UNEP/CDB/SBSTTA/18/17)15. Nos últimos 15 anos diversos estudos mostraram o efeito de diluição da biodiversidade na dispersão de agentes patogênicos e na modulação de sua dinâmica de sua transmissão [13, 26, 18]. No entanto, os estudos e ações no último século, apesar da expansão do conhecimento epidemiológico, reagiram a eventos de emergência de doenças específicas na população humana, com algumas tentativas de mitigação. Considerando a baixa capacidade de reverter as mudanças climáticas e os impactos ambientais determinados pelo crescimento humano, além de nossa forma de produ-

[1]Fundação Oswaldo Cruz
Programa Institucional Biodiversidade de Saúde (Fiocruz – PIBS)
Rio de Janeiro – RJ – Brasil
[2]Laboratório Nacional de Computação Científica (LNCC/MCTI)
Petrópolis – RJ – Brasil
mchame@fiocruz.br, hcbm@lncc.br, lgadelha@lncc.br
daa@fiocruz.br, krempser@lncc.br, abdallalivia@fiocruz.br

[15]http://www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-17-en.pdf

ção e consumo de recursos naturais, parece razoável prever que não conseguiremos deter a emergência destas doenças. Esse quadro é paradoxal em países megadiversos, como o Brasil. Ao mesmo tempo em que a riqueza de espécies existentes traz a elas associadas também a riqueza de parasitos e, portanto, um potencial risco, é esta complexidade de espécies e de suas relações que protegem e estabiliza a dinâmica das transmissões, reduzindo o surgimento de surtos de doenças, um dos mais importantes serviços ecossistêmicos. Diante deste cenário, mais que buscar respostas eficientes para situações de crise, há motivos para se buscar ações que antecipem problemas para que se possa mitigá-los quando possível, e responder rapidamente a eles quando prevenção e/ou mitigação falharem.

Essa abordagem vem sendo fortalecida com programas internacionais, como o "One world, one health" da OMS/OIE16 e o Plano Estratégico 2011-2020 da Convenção da Diversidade Biológica (CDB)17 e, estrategicamente, em programas governamentais de países desenvolvidos que já envidam recursos e esforços consideráveis para o rastreamento de patógenos em todo o mundo, quer seja para prevenção de pandemias, como as recém ocorridas com as novas gripes e Ebola, o desenvolvimento de novos fármacos ou mesmo por preocupações de guerra biológica. No Brasil são incipientes as estratégias sistematizadas para o monitoramento e a previsão de ocorrências de doenças advindas da biodiversidade, que seguem modelo de notificação de agravos já ocorridos em humanos, insuficientes para ações preventivas [2].

2. ESCOPO DA SOLUÇÃO PROPOSTA

As relações que unem a biodiversidade à saúde são complexas porque frequentemente são indiretas, dispersas no espaço e no tempo e dependentes de inúmeras forças [19]. Não se trata somente de identificar espécies e sua distribuição geográfica. No contexto da emergência de zoonoses estão imbricadas diversas espécies de patógenos, vetores e hospedeiros que modulam evolutivamente entre si, dinâmicas e composição populacional que reagem às mudanças ambientais [13].

Enfrenta-se, portanto, um desafio de múltiplas dimensões. A primeira é sensibilizar os tomadores de decisão sobre a necessidade de monitorar a circulação de patógenos na fauna silvestre antes que estes acometam humanos, ampliando as ações da vigilância em saúde para além dos humanos. A segunda dimensão é construir mecanismo que não se limite

[16]http://www.oneworldonehealth.org/
[17]http://www.cbd.int/sp/

frente ao tamanho territorial do Brasil, às políticas setoriais pouco integradas, às urgências nacionais que absorvem pessoal e não se ocupariam das tarefas do monitoramento. A terceira é como integrar múltiplas competências, já que esse mecanismo deverá abarcar especialistas para lidar com dados, espécies e contextos sociais e ambientais distintos. A quarta é como efetivamente obter informações e tratá-las adequadamente. A quinta é extrair dos dados informações relevantes e identificar realmente os riscos e prevê-los e, por fim, o compromisso de levar informações relevantes à sociedade.

Como evidenciado, a coleta de dados, o monitoramento e a extração de conhecimento e informações sobre a saúde silvestre e suas relações com a saúde humana mostram-se tarefas desafiadoras envolvendo inúmeras áreas do conhecimento, as caracterizando como atividades interdisciplinares que visam à modelagem de um sistema dinâmico e complexo.

É também evidente que grandes áreas da computação são de indispensável aplicação no contexto apresentado, tais como a modelagem computacional, a aprendizagem de máquina e a programação paralela, porém suas aplicações não são óbvias, dada à necessidade de integração de informações de diferentes meios, a complexidade e dimensionalidade dos dados a serem manipulados e a sensibilidade envolvida na utilização e divulgação desses dados.

Na construção de um sistema capaz de tratar todas as questões elencadas, elaborou-se a colaboração entre a Fundação Oswaldo Cruz (Fiocruz) e o Laboratório Nacional de Computação Científica (LNCC).

Considerando as tecnologias disponíveis e a sinergia estabelecida entre as duas instituições, o desenvolvimento do Sistema de Informação em Saúde Silvestre – SISS-Geo18 é a proposta para avançar sobre os desafios postos. Sua concepção busca a integração e a participação de diversos segmentos da sociedade, desde o registro de dados primários por qualquer pessoa interessada, na aplicação do conceito de ciência cidadã, ao diagnóstico confiável de agentes patogênicos que circulam na fauna silvestre com potencial de acometimento humano com a participação de rede de laboratórios e especialistas, até os desafios computacionais e matemáticos que incluem sistemas analíticos e de predição; mineração de dados; processos intensivos; programação paralela; integração de sistemas, dados (desestruturados e heterogêneos) e informações; geoprocessamento; aprendizagem de máquina, meta-heurísticas e visualização de dados para a construção de modelos de alerta e previsão de agravos advindos da biodiversidade e promovidos pelas forças motrizes antropogênicas.

[18]http://www.biodiversidade.ciss.fiocruz.br/apresentacao-0
[19]http://www.biodiversidade.ciss.fiocruz.br

O SISS-Geo tem como característica essencial o tratamento dos seus dados em ambiente espacialmente referenciado. Tem como objetivos: (i) proporcionar, de maneira rápida e eficiente, o fluxo de informações entre o Centro de Informação em Saúde Silvestre19 da Fiocruz e o sistema nacional de vigilância em saúde, com contribuição especial ao Centro de Informações Estratégicas em Vigilância em Saúde – CIEVS/MS; as redes participativas em saúde silvestre e de laboratórios; a população em geral que deseja participar do processo e; os diferentes centros de monitoramento da biodiversidade, como o MCTI (Ministério da Ciência, Tecnologia e Inovação), ICMBio (Instituto Chico Mendes de Biodiversidade), JBRJ (Jardim Botânico do Rio de Janeiro), MAPA (Ministério da Agricultura, Pecuária e Abastecimento), Embrapa (Empresa Brasileira de Pesquisa Agropecuária), etc.; (ii) criar, a partir dos dados e informações georreferenciadas, modelos de alerta e previsão de agravos à saúde silvestre e humana, de modo a atuar como sistema sentinela para doenças emergentes e reemergentes e, ainda, disponibilizar os resultados das modelagens espaciais para a comunidade científica e tomadores de decisão; (iii) permitir meios adequados para integração do sistema georreferenciado com bancos de dados geográficos de parceiros governamentais e não governamentais; (iv) adequar-se ao padrão de metadados da Infraestrutura Nacional de Dados Espaciais (INDE)20, visando disponibilizar, com eficiência e total compatibilidade, dados relacionados à saúde silvestre para a comunidade científica e a população em geral.

SISS-Geo está construído sobre quatro macro-módulos. O primeiro sistematiza a captação dos registros georreferenciados das observações de campo de animais e de suas condições físicas e do ambiente ao seu redor, feita por colaboradores por meio de dispositivos móveis (Android e IOS) e em ambiente Web, os organizando em bancos de dados (Seção 3.2). O segundo gera, utilizando-se da modelagem a partir de dados, modelos automatizados de alertas considerando as distâncias territoriais, os intervalos de tempo entre elas, a similaridade dos grupos taxonômicos envolvidos, com notabilidade para primatas, quirópteros, roedores e carnívoros, mas não a eles limitados, as condições físicas de encontro dos animais no campo, de acordo com padrões clínicos pré-categorizados, além das informações ambientais do local onde o animal foi avistado (Seção 3.3.1).

A partir da indicação de importância e emergência gerada pelo modelo de alerta, busca-se a integração de atores da Rede Participativa em Saúde Silvestre e, especialmente, da Rede de Laboratórios em Saúde Silvestre, e dos serviços em saúde e ambiental instituídos no País, para a coleta de amostras biológicas em animais no campo e a confiabilidade do diagnóstico.

[20]http://www.inde.gov.br

O diagnóstico confiável realimentará e validará o modelo de alerta e, a partir da correlação inicial das condições ambientais de ocorrência, espera-se estudos e geração de modelos de previsão de oportunidades ecológicas para a ocorrência de doenças oriundas da biodiversidade, o que constitui o terceiro módulo (Seção 3.3.2).

Finalmente, o quarto módulo contempla o desafio do entendimento das relações que governam o fenômeno em questão, a partir dos modelos treinados. Neste contexto, a extração de conhecimento atua como principal mecanismo de sugestão de hipóteses para posterior investigação/validação do especialista (Seção 3.3.3).

A automatização da busca de padrões de ocorrência visa tornar possível e eficiente a abrangência de informações das pessoas mais simples até especialistas e em todo o território nacional, gerar conhecimento sobre o entendimento de padrões possíveis e parâmetros que contribuem para a ocorrência de doenças, a formação, em médio e longo prazo, de pesquisadores capazes de desenvolver modelagens complexas na área da ecologia das doenças e sua gestão integrada à tecnologia de informação geográfica, e obviamente, gerar dados para a política nacional de saúde e de conservação da biodiversidade.

A proposta inspirou-se no desejo de tornar público e buscar reforços para uma longa caminhada que congrega pesquisadores, especialistas de múltiplas áreas e a sociedade para que, por meio da computação, a informação e ações de prevenção de doenças cheguem as regiões mais remotas do País. Surge da prática de muitos anos de pesquisa de campo, no semiárido brasileiro, onde informações relevantes de agravos em animais silvestres foram perdidas ou dispersas e a falta de sua sistematização impossibilitou ações importantes tanto para a contenção de doenças em humanos, quanto para a conservação das espécies.

O SISS-Geo nasce dos esforços em criar ações inovadoras e integradas para a transversalização da biodiversidade nos setores do País. Integra-se às ações da Fundação Oswaldo Cruz no "Projeto de Ações Público-privadas para a Biodiversidade" – PROBIOII[21], coordenado pelo MMA, e desenvolvido pelo FUNBIO, Embrapa, MAPA, MS, MCTI, Jardim Botânico do Rio de Janeiro, ICMBio e Fiocruz. O LNCC se juntou ao projeto da Fiocruz e garantiu sua realização numa parceria de construção de conhecimento de longo prazo.

Ações correlatas como a Rede de Laboratórios em Saúde Silvestre, que conta com 43 laboratórios nas diversas regiões do Brasil e da Rede Participativa, com mais de 1000 seguidores no Facebook e a realização da 1ª Conferência Brasileira em Saúde Silvestre e Humana [2] fundamentam o escopo da solução proposta.

[21]http://www.mma.gov.br/biodiversidade/projetos-sobre-a-biodiversidade/projeto-nacional-de-acoes-publico-privadas-para-biodiversidade-probio-ii

3. DESAFIOS E PROPOSTAS RELATIVAS AO SISS-GEO

Os principais desafios computacionais encontrados no SISS-Geo podem ser categorizados em quatro classes: gerenciamento de dados sobre saúde silvestre, geoprocessamento, aprendizagem de máquina e ferramentas de apoio à rastreabilidade e à composição de análises sobre saúde silvestre, detalhados a seguir.

3.1 Gerenciamento de Dados sobre Saúde Silvestre

Para monitorar as mudanças na biodiversidade é essencial coletar, documentar, armazenar e analisar indicadores sobre a distribuição espaço-temporal das espécies, além de obter informações sobre como elas interagem entre si e com o ambiente em que vivem [15]. O desenvolvimento e implantação de mecanismos para produzir esses indicadores dependem do acesso a dados confiáveis obtidos em expedições de campo, por sensores automáticos, em coleções biológicas e na literatura acadêmica. Esses dados normalmente estão disponíveis em diversas instituições que utilizam formatos e identificadores distintos, o que torna desafiador o trabalho de integração de dados.

As metodologias e técnicas usadas para gerenciar e analisar esses dados definem uma área de pesquisa frequentemente chamada de Informática na Biodiversidade [23, 11]. Algumas iniciativas para o estabelecimento de padrões de metadados e de publicação de dados, como o EML [7] e o Darwin Core [25], conseguiram estabelecer conjuntos de identificadores para descrever os principiais conceitos sobre biodiversidade. Embora esses identificadores cubram apenas uma fração dos conceitos possíveis, eles permitem que instituições publiquem seus dados sobre biodiversidade utilizando o mesmo formato e que estes sejam coletados e processados automaticamente por sistemas agregadores, como o GBIF22.

Por meio da utilização desses padrões, o SISS-Geo poderia coletar dados de ocorrências de espécies disponibilizados por diversos provedores, assim como oferecer os dados armazenados no seu próprio banco de dados para a comunidade em geral em um formato de fácil consumo. O Darwin Core tem sido estendido para inclusão de conceitos sobre temas específicos, como informações sobre interações e polinizadores (Interaction Extension to Darwin Core) e sobre fichas de espécies (Plinian Core). Seria importante avaliar e propor uma extensão do padrão para contemplar informações sobre saúde silvestre nos registros de observação de espécies, o que normalmente é realizado no contexto do TDWG23.

22http://www.gbif.org
23http://www.tdwg.org

3.2 Geoprocessamento

A espacialização e visualização geográfica são hoje condições básicas para a gestão da informação. Quase nunca ela é simples por questões que incluem a necessidade de normalização, atualização e acesso a dados qualificados. Nos estudos das doenças infecciosas a espacialização dos dados ainda precisa considerar pulsos e flutuações populacionais determinados por diversos fatores como sazonalidade, períodos reprodutivos, migrações, entre outros [16].

O Sistema de Informação em Saúde Silvestre – SISS-Geo tem por objetivo construir informações relevantes e confiáveis, capazes de cooperar nos processos decisórios do Ministério da Saúde, do Ministério da Agricultura, Pecuária e Abastecimento e do Ministério do Meio Ambiente, fornecendo subsídios para tomadas de decisão mais ágeis e oportunas.

Por se tratar de um projeto inovador, as tarefas desenvolvidas não são simples e não existem soluções prontas. É, portanto, necessária, a construção de novas metodologias e a utilização de diferentes tipos de tecnologias geográficas capazes de atender as expectativas e objetivos do SISS-Geo. A Infraestrutura de Geoprocessamento (IG) do SISS-Geo tem importância estratégica nesse processo, existindo a necessidade de superar desafios pertinentes ao controle da qualidade de dados geográficos, minimização dos erros posicionais dos modelos, espacialização da modelagem baseada em aprendizagem de máquina e disponibilização dos modelos sob a forma de mapas dinâmicos na Internet.

A modelagem de oportunidade ecológica de doenças do SISS-Geo irá utilizar uma densa massa de dados geográficos com escalas, sistemas de referência, fontes e metodologias de mapeamento distintas. Para isso é necessária a normalização e integração dos dados em banco de dados geográfico. Este será utilizado tanto no consumo de informações/dados, quanto no armazenamento dos resultados pertinentes à modelagem, sob a forma de modelos geograficamente distribuídos. Os dados de entrada para a modelagem são obtidos em função da sobreposição dos registros de animais silvestres com as bases de dados ambientais, sociais e de impactos antrópicos. Em função da localização dos registros, serão estabelecidos relacionamentos espaciais do tipo “está dentro”, “está próxima”, “intersecta”, etc.

As bases de dados sistemáticas disponibilizadas por fontes oficiais do governo Federal, Estadual e Municipal, são produzidas, em sua maioria, em pequena e média escala (1:1.000.000, 1:500.000). O mapeamento nestas escalas proporciona somente a visão geral do espaço, com grau de detalhamento e precisão reduzidos. Isso pode influenciar significativamente na modelagem do SISS-Geo, pois implicará em grau de incerteza entre o conjunto de pontos registrados e os dados cartográficos nacionais, à medida que estes forem relacionados.

A medida de incerteza geralmente corresponde ao Padrão de Exatidão Cartográfico (PEC), cujo valor é estimado para cada mapeamento e define a classificação de uma carta. No entanto, o uso da PEC é questionável quando se trata de cartografia digital [21], cujo desenvolvimento introduziu novas técnicas de mapeamento e de cálculos de erros. A Especificação Técnica para a Aquisição de Dados Geoespaciais Vetoriais (ET-ADGV) adotada na Infraestrutura Nacional de Dados Espaciais (INDE), também aborda essa questão e define novos parâmetros a serem seguidos em relação ao mapeamento sistemático do Brasil. Essa norma considera que a exatidão na aquisição do dado é igual a do produto cartográfico digital final, porque após a aquisição vetorial de um elemento qualquer, sua geometria não é mais alterada nos processos posteriores. Além disso, os padrões de exatidão considerados nessa norma são bem mais rigorosos que os baseados em cartografia analógica e são calculados com base na comparação estatística entre medições realizadas em campo e no produto digital. A adoção ET-ADGV é uma tendência, mas ainda está sob processo de adaptação, de modo que poucos dados terão essa informação documentada. Portanto, a referência do valor de exatidão posicional dos dados para as consultas espaciais do SISS-Geo será inicialmente baseada na PEC.

Para a minimização do efeito do erro posicional nos modelos do SISS-Geo será considerado a tolerância nos cruzamentos espaciais, com base na exatidão posicional dos dados em sobreposição, utilizando como referência a PEC. Busca-se com isso, estabelecer modelos com qualidade posicional suficiente para apoiar as tomadas de decisão oriundas de políticas de saúde pública.

A infraestrututra de geoprocessamento necessita também disponibilizar, ao domínio público24, os resultados e os modelos de alerta e previsão do SISS-Geo, excetuando-se as informações sensíveis25. Portanto, segue em desenvolvimento a adequação do sistema de informação geográfica para ambiente Web, que irá disponibilizar os resultados do SISS-Geo sob a forma de mapas dinâmicos/interativos e estatísticas gráficas na Internet. Como vantagem dessa tecnologia, encontra-se a facilidade de manipulação, análise e interpretação dos modelos pelo usuário final; independência de sistema operacional e interação com sistemas desktop ou outros sistemas da Internet (interoperabilidade).

3.3 Aprendizagem de Máquina

3.3.1 Agrupamento de registros de observações e predição de alerta

Quando a observação de um animal silvestre, sua condição física e ambiente circundante é registrada no SISS-Geo, seja por especialistas ou colaboradores do sistema,

[24]Lei de acesso à informação: http://cidadao.mpf.mp.br/acesso-a-informacao
[25]Instrução Normativa da INDA: http://dados.gov.br/instrucao-normativa-da-inda

esta é reunida com outros registros relacionados (previamente comunicadas) dando origem a coleção de acontecimentos que caracterizam um fenômeno. Esta é a fase de agrupamento e, embora possa soar descomplicada, envolve o desafio de se conceber/treinar modelos dotados de capacidade discriminativa em reconhecer similaridades e dissimilaridades entre eventos, baseando-se em critérios como distância espacial e temporal entre os registros, similaridade entre espécies e nas condições físicas reportadas e outros. Este fluxo de aprendizagem é sintetizado na **Figura 1**.

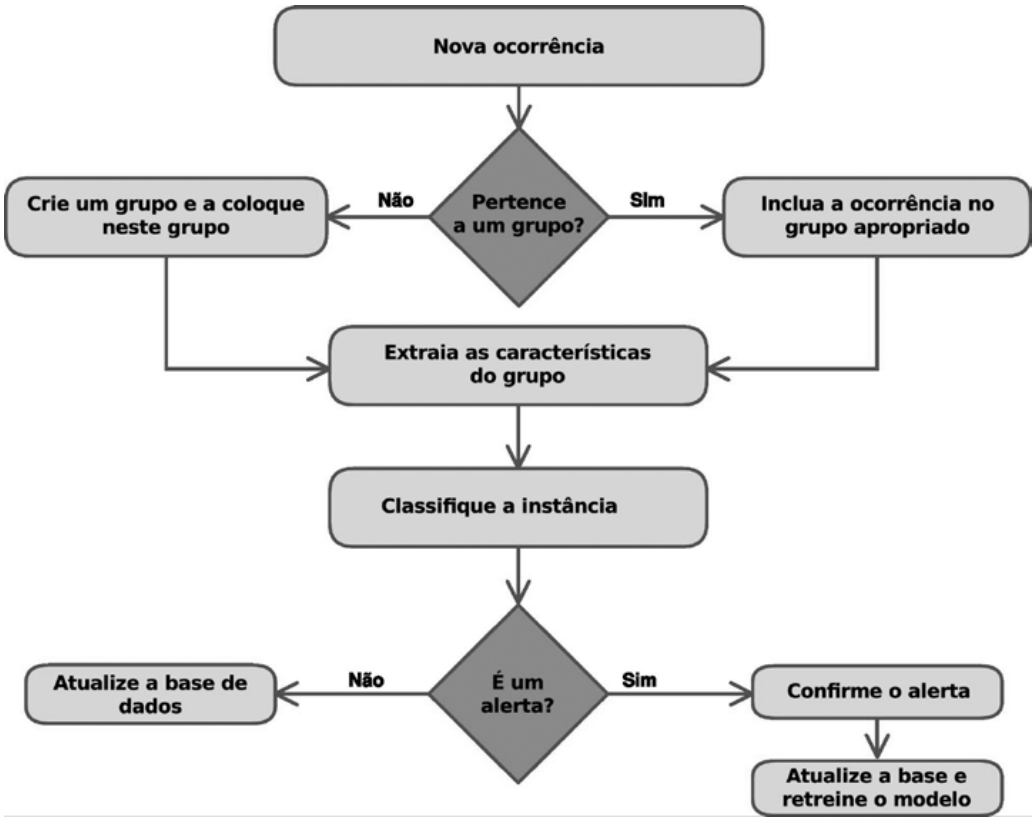

Figura 1: Fluxo relativo à aprendizagem de máquina do SISS-Geo.

A segunda parte consiste em modelar características dos registros das observações que as tornam menos ou mais relevantes, isto é, treinar o modelo de alertas. Significa predizer a gravidade dos registros de acordo com as informações trazidas pelos eventos bem como o contexto geográfico/ambiental. Por exemplo, um registro envolvendo isoladamente um animal com sintomas é em geral menos grave do que ocorrências contendo eventos similares, mas abrangendo grupos de animais. Naturalmente, em situações reais a caracterização de uma situação de alerta costuma ser bem menos óbvia, usualmente considerando vários fatores para a tomada de decisão.

Percebe-se que as atividades acima mencionadas referem-se à tarefa de agrupamento e classificação de dados, típicas da aprendizagem de máquina, e notoriamente conhecidas pela ampla pluralidade de abordagens e metodologias. São assim tarefas complexas, tanto pela natureza como também pelo grande volume de dados esperado para o sistema26.

No entanto, os desafios do agrupamento e da classificação que se manifestam no SISS-Geo vão além dos desafios clássicos destas tarefas.

Caracterização de um fenômeno.

A caracterização do que define um grupo de ocorrências (fenômeno) recai no problema da formulação de medidas de similaridade não convencionais (p.e., não necessariamente Euclidianas). Regras de agrupamento baseadas na experiência do especialista constituem uma alternativa razoável, mas esbarram na limitação da formalização do conhecimento e consequente potencial de introdução de vieses indesejáveis. Uma outra abordagem é tratar este problema como um processo de aprendizagem de máquina, objetivando o treinamento de modelos de similaridade: dado um novo registro e o conjunto de registrados existentes, determinar a qual grupo ele pertence—ou se caracteriza um novo grupo. O processo configura-se como aprendizado supervisionado, uma vez que é possível determinar confiavelmente, a priori ou a posteriori, quais registros pertencem a quais fenômenos, seja por exames laboratoriais ou convicção de especialistas.

Extração de características.

Uma vez constituídos os fenômenos é preciso avaliá-los quanto às potencialidades de ameaça à saúde silvestre e possível acometimento a humanos, pois, fenômenos por si só não configuram situações de alerta. Neste sentido, informações que caracterizam um grupo de ocorrências precisam ser extraídas e fornecidas ao modelo de predição de alerta. A dificuldade é, assim, derivar quais estatísticas melhor representam o fenômeno descrito pelo grupo a fim de maximizar o desempenho do modelo de predição; em outras palavras, levantar as informações que facilitem o processo de aprendizado. Especialistas preconizam o uso de certas estatísticas, como a espécie e quantidade de animais acometidos, número e frequência das ocorrências, entre outras; no entanto, o espaço de possíveis características vai muito além e poderia-se melhorar o desempenho preditivo. Dessa forma, uma questão em aberto é: como explorar esse vasto espaço automaticamente? Uma interessante linha de pesquisa e potencial solução para este desafio é a investigação de métodos de extração automática de características [10, 9].

[26]Afinal, é um sistema ambicioso que almeja agregar e hospedar os registros sobre saúde silvestre do vasto território nacional.

Modelo de predição de alerta.

Embora o seu uso no sistema aproxime-se de métodos suficientemente conhecidos e descritos na literatura, o modelo de predição de alerta é provavelmente o componente mais estratégico da inteligência do SISS-Geo. A viabilidade do sistema está fundamentalmente calçada no desempenho em termos de acurácia do modelo de predição, tanto na detecção de verdadeiro positivos (alertas) como de verdadeiro negativos (não alertas). A não detecção de uma situação de alerta (falso negativo) pode resultar em consequências graves à saúde silvestre, ambiental e, também, à humana. Por outro lado, os falsos positivos sobrecarregariam a escassa rede de laboratórios e especialistas responsáveis por confirmar ou negativar alertas (mais detalhes a seguir). Nesse sentido, métodos que combinam diversos modelos (ensemble methods) usualmente produzem soluções mais acuradas e robustas, sendo, portanto, candidatos promissores como algoritmos de treinamento dos modelos de predição [20]. Ainda, uma vez que a grande parcela de dados do sistema não possui classe associada, isto é, os fenômenos cujas predições de alerta ainda não foram confirmadas, o aprendizado semi-supervisionado constitui uma abordagem interessante em função da capacidade em também aproveitar instâncias não classificadas no processo de treinamento [3].

Confirmação de alerta.

Outro componente-chave do SISS-Geo—e do qual todos os demais dependem—é o processo de confirmação de alertas. O grande desafio e gargalo decorrem da necessidade da participação direta de humanos no procedimento de confirmação, seja em campo ou laboratório; portanto, é um processo caro e lento, mesmo considerando a extensa rede de colaborações qualificadas ligadas ao SISS-Geo. Quando há mais alertas emitidos pelo modelo de predição do que a capacidade de especialistas e rede de laboratórios em confirmá-los, os fenômenos precisam ser priorizados. Nesta situação, pode-se pensar em priorizar os fenômenos associados a alertas (1) pelo nível de alerta ponderado pela confiança da predição; ou (2) pela pertinência às regiões de grande interesse, seja este social, ambiental e/ou econômico. Entretanto, uma estratégia com enfoque em médio e longo prazo é a priorização da confirmação (ou negação) de alertas com maior potencial de aprimoramento da acurácia do modelo de predição. Esta linha de pesquisa é recente e denominada active learning [22]. Este mesmo método pode ser também empregado nos eventuais casos de falso negativos, evitando-se assim a possibilidade de degeneração do modelo de predição27: os fenômenos preditos como não alertas mas promissores sob o ponto de vista de aprendizado seriam passíveis de confirmação (da condição de não alerta) por especialista.

[27]Considere a situação extrema em que todas as predições são de *não alertas*, incluindo tanto verdadeiros quanto falso negativos. Dado que em princípio somente os casos de alertas são de interesse e passíveis de confirmação, neste cenário o modelo estaria fadado à degeneração.

3.3.2 Previsão de oportunidades ecológicas de ocorrência de doenças

Outra linha de fundamental importância no SISS-Geo é a previsão de cenários e ambientes que favoreçam oportunidades ecológicas para a ocorrência de doenças advindas da fauna silvestre ou, posto diferentemente, o levantamento de cenários propícios à ocorrência de certo evento, como por exemplo, um surto de uma determinada doença.

Resumidamente, os modelos de alerta treinados podem ser empregados para a avaliação de diversos cenários e caracterizar aqueles potencialmente suscetíveis. A construção dos modelos de previsão deverá relacionar diversas informações ambientais, sociais e de saúde humana e animal, mostrando-se uma área desafiadora aos atuais modelos de previsão. Métodos de relacionamento de variáveis ambientais e animais, tais como os aplicados na modelagem de nichos ecológicos ou mesmo os mais tradicionais métodos de aprendizagem de máquina serão amplamente aplicados nesse contexto, porém, novas abordagens devem ser elaboradas, permitindo a integração da variedade de informações citadas.

Além disso, devem-se considerar os desafios computacionais envolvidos na manipulação de informações de um grande número de registros, de diferentes espécies e condições ambientais, definindo-se como um problema de alto custo computacional. Entretanto, apesar da esperada manipulação de grandes massas de dados, espera-se também um reduzido número de informações sobre uma espécie ou doença específica, levando a um novo desafio: a aplicação de técnicas de predição em um ambiente altamente desbalanceado.

3.3.3 Extração de conhecimento

Uma propriedade importante dos métodos de modelagem simbólica, como árvores de decisão, algoritmos de extração de regras e a meta-heurística programação genética [14], é que o modelo é, ele próprio, a representação explícita do conhecimento extraído dos dados. Mais especificamente, é a revelação—passível de interpretação humana—das relações existentes entre os dados de entrada e saída.

É notável a potencialidade desta classe de modelos em assistir especialistas na análise e entendimento do fenômeno investigado, levando a interação homem-máquina curiosa: o modelo sugere hipóteses que melhor se ajustem aos dados enquanto o especialista as valida.

O desafio central da extração de conhecimento está na definição da estrutura/linguagem do modelo ou, em outras palavras, na incorporação do conhecimento do especialista. Nesse sentido, é também desafiador encontrar o balanço ideal entre viés, geralmente decorrente da simplicidade estrutural do modelo, e variância, questão normalmente associada aos modelos estruturalmente mais complexos.

Dependendo da parametrização dos algoritmos de aprendizagem e dimensionalidade da base de dados, um segundo desafio emerge: a demanda computacional associada, que é agravada pelo fato do processo de extração de conhecimento ser muitas vezes realizado interativamente pelo especialista. Tipicamente, as estratégias empregadas nestas situações incluem a (1) redução da dimensionalidade dos dados [8] e (2) computação paralela e distribuída, em arquiteturas convencionais ou aceleradores [1].

3.4 Ferramentas de Apoio à Rastreabilidade e à Composição de Análises sobre Saúde Silvestre

Ferramentas de análise e síntese de dados de biodiversidade a exemplo da modelagem de distribuição de espécies (MDE) [24], são amplamente utilizadas. Essas análises normalmente empregam diversas aplicações distintas, executadas de forma fracamente acoplada, caso típico para a utilização de sistemas de gerenciamento de workflows científicos [5]. Por exemplo, no caso da MDE, dados ambientais globais, como climatologia, uso da terra e topografia, são recuperados de provedores de dados ambientais, enquanto dados de ocorrências de espécies são obtidos de provedores como o GBIF. É comum que esses dados tenham que ser adaptados com sistemas de informações geográficas ou filtrados com ferramentas de controle de qualidade. Após esses passos de pré-processamento, algoritmos para MDE, a exemplo do Maxent [17], são aplicados para prever a distribuição potencial de espécies utilizando os dados ambientais e os de ocorrência de espécies adquiridos e manipulados nos passos de pré-processamento. Finalmente, uma etapa de pós-processamento é realizada, onde ferramentas estatísticas e de visualização de dados são utilizadas para analisar os dados resultantes da modelagem.

A utilização de sistemas de gerenciamento de workflows científicos permite que tais composições de diversas atividades, como ferramentas, sejam mais fáceis de especificar e executar por meio da automação de rotinas que frequentemente estão envolvidas nas mesmas:

- as atividades podem ter dependências de dados entre si, de modo que algumas atividades só podem iniciar a sua execução quando seus dados de entrada, produzidos por outras atividades que as precedem, estiverem disponíveis;

- eventualmente o fluxo de execução do workflow pode sofrer uma bifurcação para a execução de diversas atividades independentes entre si, tornando interessante a execução paralela das mesmas por questão de escalabilidade;

- o inicio da execução de uma atividade pode depender do fim da execução de diversas atividades disparadas após uma bifurcação, requerendo uma sincronização da execução do workflow, onde é necessário garantir que todas as atividades geradas pela bifurcação de fato terminaram a sua execução;

• caso o workflow utilize diversos recursos computacionais remotos, é necessário gerenciar a transferência de dados e monitorar a execução de atividades remotas.

Informações de proveniência [4], que reúnem detalhes sobre a concepção e a execução de processos computacionais, como por exemplo, workflows científicos, descrevendo os processos e dados envolvidos na geração dos resultados dos mesmos, podem ser utilizadas para facilitar esta tarefa. Elas permitem a descrição precisa de como o processo computacional foi projetado, chamada de proveniência prospectiva, e do que ocorreu durante sua execução, denominada proveniência retrospectiva. Algumas aplicações da proveniência incluem a reprodução de um processo computacional para fins de validação, compartilhamento e reutilização de conhecimento, verificação de qualidade de dados e atribuição de resultados científicos. Um dos conceitos comumente capturados na proveniência é o de causalidade, que é dado pelas relações de dependência existentes entre atividades computacionais e conjuntos de dados. Estas dependências podem derivar, por transitividade, dependências entre conjuntos de dados e entre processos. No contexto do SISS-Geo, as informações de proveniência permitirão que o processo de geração de alertas seja rastreável, ou seja, que seja possível recuperar os dados, parâmetros de configuração e atividades computacionais utilizados.

REFERÊNCIAS

[1] Augusto, D. A.; Barbosa, H. J. C. Accelerated parallel genetic programming tree evaluation with OpenCL. Journal of Parallel and Distributed Computing, Volume 73, Issue 1, Pages 86-100, 2012.

[2] Chame, M; Labarthe, N. Saúde Silvestre e Humana: Experiências e perspectivas. Fundação Oswaldo Cruz, Fiocruz, Rio de Janeiro, 2013.105p, 2013.

[3] Chapelle, O.; Schölkopf, B.; Zien, A. Semi-supervised learning. Cambridge, Mass.: MIT Press, 2006.

[4] Cuevas-Vicenttín, V.; Dey, S.; Köhler, S.; Riddle, S.; and Ludäscher, B. Scientific Workflows and Provenance: Introduction and Research Opportunities. Datenbank-Spektrum, 12(3):193–203, 2012.

[5] Deelman, E.; Gannon, D; Shields, M.; Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25(5):528–540, 2009.

[6] Estrada-Peña, A.; Ostfeld, R. S.; Peterson, A. T; Poulin, R.; Fuente, J. Effects of Environmental change on zoonotic disease risk: an ecological primer. Trends in Parasitology, 30(4):205-214, 2014.

[7] Fegraus, E. H.; Andelman, S; Jones, M. B.; Schildhauer, M. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. Bulletin of the Ecological Society of America, 86(3):158–168, 2005.

[8] Fodor, I. A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National, Technical Report, 2002.

[9] Guo, L.; Rivero, D.; Dorado, J.; Munteanu, C. R.; Pazos, A. Automatic feature extraction using genetic programming: An application to epileptic EEG classification. Expert Systems with Applications, 2011.

[10] Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. Feature Extraction, Foundations and Applications. Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[11] Hobern, D.; Apostolico, A.; Arnaud, E.; Bello, J. C.; Canhos, D.; Dubois, G.; Field, D.; García, E.; Hardisty, A.; Harrison, J.; Heidorn, B.; Krishtalka, L.; Mata, E.; Page, R.; Parr, C.; Price, J.; Willoughby, S. Global Biodiversity Information Outlook - Delivering Biodiversity Knowledge in the Information Age. Technical report, GBIF Secretariat, 2013.

[12] Jones, K.; Patel, N. G.; Levy, M. A.; Storeygard, A.; Balk, D.; Gitleman, J. L.; Daszak, P. Global trends in emerging infectious diseases. Nature, 451:990-993, 2008.

[13] Keesing, F.; Holt, R. D; Ostfeld, R. S. Effects of species diversity on disease risk. Ecology Letters, 9:485-495, 2006.

[14] Koza, J. R. Genetic programming: On the programming of computers by natural selection. MIT Press, Cambridge, Mass., 1992.

[15] Michener, W. K; Jones, M. B. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution, 27(2):85–93, 2012.

[16] Ostfeld, R; Glan, G. E.; Keesing, F. Spatial epidemiology: an emerging (or-reemerging) discipline. Trends in Ecology and Evolution, 20(6): 328-336, 2006.

[17] Phillips, S. J.; Anderson, R. P.; Schapire, R. E. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3-4):231–259, 2006.

[18] Poulin, R.; Forbes, M. Meta-analysis and research on host-parasite interactions: past and future. Evol. Ecol., 26:1169-1185, 2012.

[19] Poulin, R. Network analysis shinning light on parasite ecology and diversity. Trends in parasitology, 26:492-498, 2010.

[20] Rokach, L. Ensemble-based classifier. Artificial Intelligence Review, Volume 33, Issue 1-2, pp 1-39, 2010.

[21] Santos, S. D. R.; Huinca, S. C. M. Considerações sobre a utilização da PEC Padrão de Exatidão Cartográfica nos dias atuais. III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias de Geoinformação. Recife, Pernambuco, 2009.

[22] Settles, B. Active Learning Literature Survey. Computer Sciences Technical Report, University of Wisconsin–Madison, 2009.

[23] Soberón, J.; Peterson, A. T. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 359(1444):689–98, 2004.

[24] Peterson, A. T.; Soberón, J.; Pearson, R. G.; Anderson, R. P.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M. B. Ecological Niches and Geographic Distributions. Princeton University Press, 2011.

[25] Wieczorek, J.; Bloom, D.; Guralnick, R.; Blum, S.; Döring, M.; Giovanni, R.; Robertson, T.; Vieglais, D. Darwin Core: an evolving community-developed biodiversity data standard. PloS One, 2012.

[26] Xavier, S.D.C.; Roque, A.L.R.; Lima, V.S.; Monteiro, K.J.L., Otaviano, J.C.R. Lower Richness of small wild mammals species and Chagas disease risk. PLoS Neglected Tropical Diseases, 2012.