UNIOESTE – Universidade Estadual do Oeste do Paraná

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

Métodos de pré-análise de dados para modelagem de distribuição de espécies

Thiago Junior Vacari

CASCAVEL

2014

THIAGO JUNIOR VACARI

MÉTODOS DE PRÉ-ANÁLISE DE DADOS PARA MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientador: Adair Santa Catarina

CASCAVEL

2014

THIAGO JUNIOR VACARI

MÉTODOS DE PRÉ-ANÁLISE DE DADOS PARA MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

Monografia apresentada como requisito parcial para obtenção do Título de *Bacharel em Ciência da Computação*, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Adair Santa Catarina (Orientador)
Colegiado de Ciência da Computação,
UNIOESTE

Prof. Dr. Claudia Brandelero Rizzi
Colegiado de Ciência da Computação,
UNIOESTE

Prof. M.Eng. Carlos José Maria Olguin
Colegiado de Ciência da Computação,
UNIOESTE

Cascavel, 05 de novembro de 2014.

AGRADECIMENTOS

A Deus, por ter me dado saúde e força para superar as dificuldades.

Ao meu professor orientador, Dr. Adair Santa Catarina, pelo auxilio, disponibilidade de tempo e material, e pelo incentivo.

A esta universidade e seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, contagiado pela acendrada confiança no mérito e ética aqui presentes.

Aos meus pais que proporcionaram a minha existência e com muito carinho e apoio, não mediram esforços para que eu chegasse a esta etapa da minha vida.

A minha namorada pelo apoio e incentivos nas horas mais difíceis.

Aos meus colegas de graduação e a todos que direta ou indiretamente colaboraram para a minha formação, o meu muito obrigado.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmo Genético
AUC	*Area Under the Curve*
FN	Falso Negativo
FP	Falso Positivo
MPG	Matriz de Proximidade Generalizada
ROC	*Receiver Operating Characteristic*
SAHGA	*Spatially Aware Hybrid Genetic Algorithm*
SDM	*Species Distribution Modelling*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

RESUMO

Sistemas para geração de Modelos de Distribuição de Espécies (SDM) têm sido apontados como ferramentas eficazes, por exemplo, no estudo da interferência humana em um habitat, na predação entre espécies e na luta pela preservação de espécies em extinção. O sistema SAHGA SDM, utilizado na modelagem de distribuição de espécies, tem como entrada variáveis ambientais e climáticas, representadas em *layers* geográficos. Estas variáveis delimitam a sobrevivência da espécie modelada. Quando são utilizadas muitas variáveis o conjunto de dados de entrada torna-se volumoso, aumentando o tempo necessário para o sistema ajustar um SDM. Visando reduzir o número de variáveis ambientais utilizadas, sem alterar a qualidade dos modelos gerados, implementou-se o processo de pré-análise de dados, através dos algoritmos Jackknife, Correlação Linear e Chi-Quadrado. Os testes realizados utilizaram a base de dados da espécie *Thalurania furcata boliviana* que contém 130 pontos de presença ou ausência da espécie e 8 variáveis climáticas. O modelo *best-subset* com todas as variáveis apresentou acurácia igual a 92,91% em um tempo igual a 9,61 minutos. O método Jackknife identificou que uma variável climática não exercia influência sobre a espécie e poderia ser suprimida na modelagem e o modelo gerado apresentou a acurácia igual a 91,98% em um tempo de 8,86 minutos. O método de Correlação Linear apontou 5 variáveis ambientais altamente correlacionadas, permitindo a exclusão de 4 delas; o modelo gerado apresentou acurácia igual a 93,27% em um tempo de 6,46 minutos. O método do Chi-Quadrado não identificou variáveis independentes e, portanto, não permitiu excluir nenhuma variável climática. Conclui-se que os métodos Jackknife e Correlação Linear podem ser usados no processo de pré-análise, identificando variáveis climáticas a serem excluídas dos modelos sem comprometer a qualidade dos mesmos.

Palavras-chave: *Layers* geográficos, Modelagem de distribuição de espécies, Seleção de variáveis.

Capítulo 1

Introdução

Cientistas utilizam sistemas para modelar como uma espécie está distribuída no espaço. Os dados de entrada utilizados nestes sistemas são pontos georreferenciados onde ocorre a presença ou ausência de exemplares da espécie, informações climáticas como temperatura, precipitação, umidade, etc., e/ou informações ambientais como, por exemplo, o tipo de vegetação ou índice de cobertura vegetal. Os dados são então processados por algoritmos computacionais e/ou estatísticos gerando como saída um modelo de distribuição de espécie (SDM) materializado num mapa da distribuição potencial da espécie para a área em estudo.

Os SDM constituem-se numa classe de ferramentas eficientes para realizar estudos do cotidiano ecológico dos seres vivos (BORGUESAN, 2013) e são fundamentados em amostras de campo, ou seja, utilizam o fato de uma espécie estar presente numa determinada área de estudo.

Um sistema para criar modelos de distribuição de espécies é o SAHGA SDM – *Spatially Aware Hybrid Genetic Algorithm for Species Distribution Modelling* – que utiliza um Algoritmo Genético (AG) com relacionamentos espaciais representados através de uma Matriz de Proximidade Generalizada (MPG) (SANTA CATARINA, 2009).

O SAHGA SDM tem a estrutura básica de um sistema para a modelagem de distribuição de espécies. Seu diferencial está no fato de inserir o conceito de dependência espacial no processo de modelagem utilizando, para tanto, a MPG para representar explicitamente o grau de associação entre os dados de entrada (pontos de presença/ausência) (SANTA CATARINA, 2009).

O algoritmo SAHGA SDM utiliza, além de pontos de presença/ausência e a MPG, um conjunto de *layers* geográficos para representar os parâmetros climáticos e ambientais que influenciam na sobrevivência da espécie (SANTA CATARINA, 2009).

Algoritmos SDM podem utilizar muitos *layers* geográficos, referentes às variáveis climáticas/ambientais, armazenadas em arquivos volumosos (RODRIGUES, 2012). Quanto

maior o número de *layers* e maior a sua complexidade, maior o processamento e o tempo necessário para o ajuste dos modelos.

Uma tentativa de acelerar o processo é realizar a pré-análise de dados. O objetivo principal deste trabalho é a implementação e os estudos dos métodos Jackknife, Correlação Linear e Chi-Quadrado para a pré-análise de dados no sistema SAHGA SDM.

Jackknife é um algoritmo para reduzir a dimensão do espaço de características, ou seja, reduzir variáveis de entrada do SDM (RODRIGUES, 2012). No SAHGA SDM esta técnica foi utilizada para estimar a importância de cada variável ambiental/climática na predição do modelo de distribuição da espécie em estudo, isto é, se o algoritmo pode reduzir o número de variáveis ambientais/climáticas utilizadas, sem afetar consideravelmente a precisão do modelo gerado.

Já o método estatístico Correlação Linear mostra o grau de relacionamento entre duas variáveis. O resultado deste método é um parâmetro que se encontra entre -1.0, passando pelo 0.0, até 1.0; este parâmetro descreve o grau de correlação entre duas variáveis ambientais (CORREA, 2003). Quando duas variáveis são fortemente correlacionadas uma delas pode ser retirada do modelo, sem prejudicar o ajuste do mesmo.

E por último, o método Chi-Quadrado, simbolizado por x^2, é um teste de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis, ou seja, informa a medida em que os valores observados se desviam do valor esperado, caso as duas variáveis não estejam correlacionadas. Quanto maior o Chi-Quadrado, mais significante é a relação entre as variáveis (LIRA, 2004).

1.1 Objetivos

O objetivo geral deste trabalho é realizar o estudo e a implementação dos métodos de pré-análise no sistema SAHGA SDM, com a perspectiva em obter quais são as *layers* geográficas importantes para o modelo e consequentemente acelerando o processo para a geração de modelos.

Para cumprir esse objetivo serão implementados três métodos de pré-análise: Jackknife, Correlação Linear e Chi-Quadrado.

Como objetivo secundário deste trabalho pode-se destacar a correção de alguns detalhes na interface do sistema SAHGA SDM a fim de torná-lo mais intuitivo e facilmente utilizável pelo usuário, bem como a correção de erros presente no sistema.

1.2 Organização do Texto

No segundo capítulo tem-se o referencial teórico, que visa apresentar os conceitos necessários à compreensão do trabalho desenvolvido. Fracionou-se o capítulo 2 em três seções. A primeira relata os conceitos referentes aos SDMs; a segunda seção aborda o processo de ajuste dos SDMs através de aplicativos de código aberto; a terceira seção descreve o sistema SAHGA SDM, que emprega o algoritmo SAHGA para modelar a distribuição potencial de espécies.

No terceiro capítulo detalham-se os métodos de pré-análise explanando seus significados e indicando seu emprego no trabalho. O quarto capítulo apresenta o estudo de caso realizado e os resultados obtidos. Conclusões e trabalhos futuros são apresentados no capítulo cinco.

Capítulo 2

Referencial Teórico

Este capítulo tem como objetivo apresentar os conceitos necessários para compreensão do trabalho desenvolvido.

A primeira seção do capítulo apresenta a estrutura dos SDMs e os mecanismos de avaliação de SDMs que serão utilizados nesse trabalho: a matriz de confusão, as curvas ROC e o índice AUC.

A segunda seção aborda o processo de ajuste dos SDMs através de aplicativos de código aberto; a terceira seção descreve o sistema SAHGA SDM, que emprega o algoritmo SAHGA para modelar a distribuição potencial de espécies.

2.1 Modelagem de Distribuição de Espécie (SDMs)

Diversos fatores levaram os pesquisadores a estudar diferentes métodos para distribuir espécies, tais como a luta pela preservação de uma espécie em extinção, o controle de espécies invasoras, a predação entre espécies e a interferência humana diminuindo os nichos potenciais (FINAMORE, 2010).

Nos últimos anos foram desenvolvidas e utilizadas várias técnicas de modelagem de distribuição geográfica de espécies com os mais variados objetivos. Algumas dessas técnicas envolvem modelagem baseada em análise ambiental, nas quais os algoritmos procuram por condições ambientais semelhantes àquelas onde as espécies foram encontradas, resultando em áreas potenciais onde as condições ambientais seriam propícias ao desenvolvimento dessas espécies (SIQUEIRA, 2005).

Os SDMs utilizam-se da modelagem matemática, aliada a ferramentas computacionais, para prever a presença ou ausência de uma espécie numa determinada área de estudo (GUISAN; THUILLER, 2005) (IWASHITA, 2007).

Os SDMs buscam definir as limitações ambientais das espécies nas dimensões para as quais o modelo é desenvolvido. Assim, informações relacionadas aos pontos de coleta podem

ser projetadas em espaço geográfico, e os algoritmos identificam locais com características ambientais similares na área de análise, indicando onde as espécies são potencialmente capazes, ou não, de manter populações viáveis (SIQUEIRA, 2005).

Esses modelos permitem identificar espécies que poderiam ser utilizadas em trabalhos de recuperação ambiental, avaliar o potencial de ameaça de espécies invasoras, avaliar o impacto das mudanças climáticas na biodiversidade, estudar possíveis rotas de disseminação de doenças infecciosas e auxiliar na determinação de áreas prioritárias para conservação.

O uso de modelagem pode indicar áreas de distribuição potencial para espécies em risco de extinção. Neste caso, o modelo resultante determina possíveis locais de ocorrência dessas espécies, aumentando a base de conhecimento a respeito da situação de risco em que se encontram. Esses resultados podem auxiliar na indicação de áreas onde existam as condições ambientais ideais para a sobrevivência das espécies analisadas, as quais podem ser recomendadas como áreas prioritárias para a conservação de espécies ou podem ser definidas como áreas potenciais para a reintrodução destas espécies (SIQUEIRA, 2005).

Fazer uso de técnicas de modelagem de distribuição geográfica de espécies é particularmente indicado em situações nas quais é preciso tomar decisões mas ainda não existe disponível um conjunto grande de informações, como é o caso de várias áreas dos principais biomas brasileiros, nas quais a coleta de dados ainda é bastante precária para o tamanho e riqueza biológica de cada região (SIQUEIRA, 2005).

Modelos de distribuição de espécies são modelos empíricos que relacionam observações de ocorrência de uma espécie com variáveis de predição de ambiente, usadas presenças estatísticas ou modelos teóricos (GUISAN; ZIMMERMAN, 2000).

Todos os estudos que envolvem SDM possuem três componentes básicos: primeiramente há um conjunto de dados que contém as variáveis explicativas; após, há um modelo matemático que relaciona a espécie com a variável explicativa; finalmente, há a avaliação da utilidade do modelo através de métodos de robustez ou por avaliação (GUISAN; ZIMMERMAN, 2000). Isto é representado na Figura 2.1.

Os sistemas para geração de SDMs estudados neste trabalho utilizam modelagem matemática juntamente com ferramentas computacionais para prever a ocorrência de determinada espécie representando-a em superfícies temáticas, indicando presença ou ausência da espécie modelada (GUISAN; THUILLER, 2005) (SANTA CATARINA, 2009). A Figura 2.2 apresenta a estrutura de um sistema para geração de SDM.

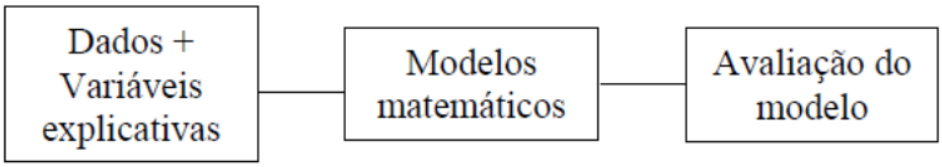

Figura 2. 1: Elementos essenciais na modelagem de distribuição de espécies

Fonte: IWASHITA (2007)

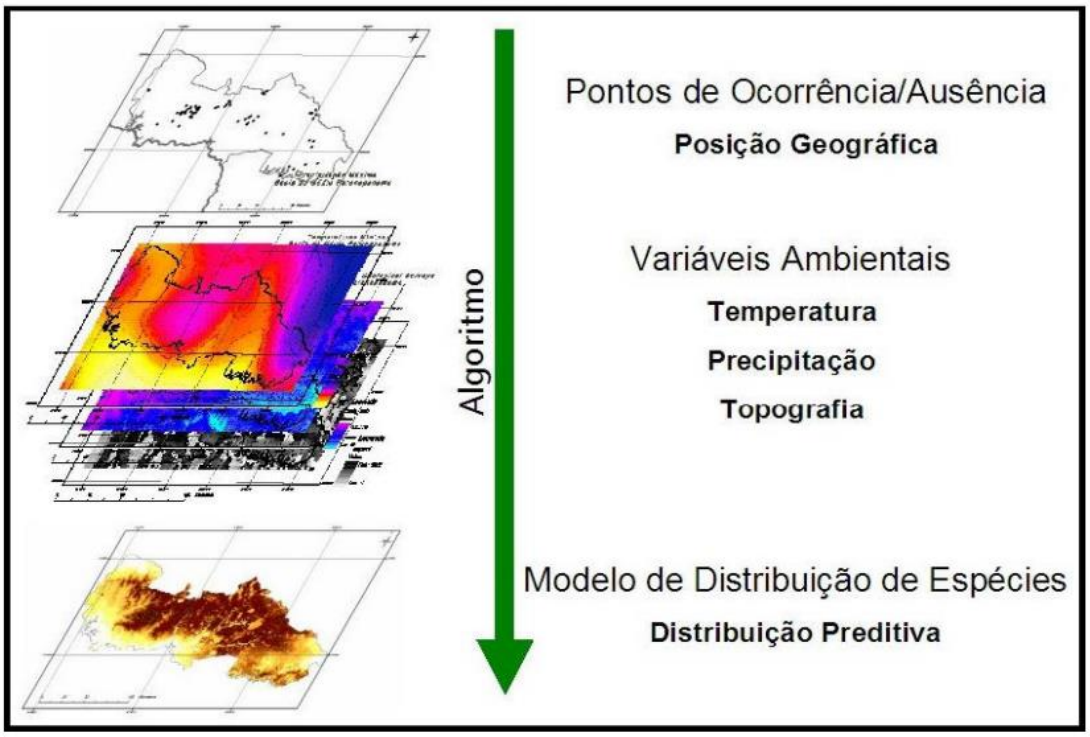

Figura 2. 2: Estrutura geral para geração de um SDM

Fonte: SIQUEIRA (2005)

Os modelos ajustados são avaliados utilizando os mecanismos apresentados na sequência: matriz de confusão, curvas ROC e o índice AUC.

2.1.1 Matriz de Confusão

Para avaliação dos modelos, uma amostra de pontos de ocorrência da espécie é classificada como presença ou ausência e comparada com os dados observados por uma matriz de confusão (NABOUT et al., 2009). O formato da matriz de confusão é apresentado na Figura 2.3.

	Amostras	
Previsão (Modelo)	Presente	Ausente
Presente	VP	FP
Ausente	FN	VN

Figura 2. 3: Matriz de Confusão

Fonte: SANTA CATARINA (2009)

Os valores VP (Verdadeiro Positivo = a) e VN (Verdadeiro Negativo = d) são predições corretas. FP (Falso Positivo = b) e FN (Falso Negativo = c) são considerados erros de predição. Os erros do tipo FP também são conhecidos como erros de comissão ou superestimativa, enquanto os erros do tipo FN são conhecidos como erros de omissão. Já o N é a quantidade total de dados de presença/ausência da espécie observada (SANTA CATARINA,2009).

A partir da matriz de confusão, podemos gerar outros valores para a avaliação dos SDMs, apresentado na tabela 2.1.

Tabela 2. 1: Medidas derivadas da matriz de confusão

Medida	Fórmula
Acurácia	$(a + d)(a + b + c + d)$
Prevalência	$(a + c) / N$
Poder de diagnóstico global	$(b + d) / N$
Taxa de classificação correta	$(a + d) / N$
Sensibilidade	$a / (a + c)$
Especificidade	$d / (b + d)$
Taxa de falso positivo (comissão)	$b / (b + d)$
Taxa de falso negativo (omissão)	$c / (a + c)$
Coeficiente de correlação de Matthews	$\frac{(a * d - b * c)}{((a + b)(a + c)(d + b)(d + c))^{1/2}}$

Fonte: SANTA CATARINA (2009)

A acurácia mede o acerto global do sistema. A sensitividade do modelo é definida como a proporção de presenças verdadeiras em relação ao total de presenças preditas pelo modelo, enquanto que a especificidade do modelo é a proporção de ausências verdadeiras em relação ao total de ausências preditas pelo modelo

2.1.2 Curvas ROC e o Índice AUC

Outro método utilizado na avaliação de SDM é a curva ROC (*Receiver Operating Characteristic*). A curva ROC é representada num gráfico de sensibilidade vs. especificidade, calculadas a partir da tabela 2.1. Exemplos de curvas ROC são apresentadas na Figura 2.4.

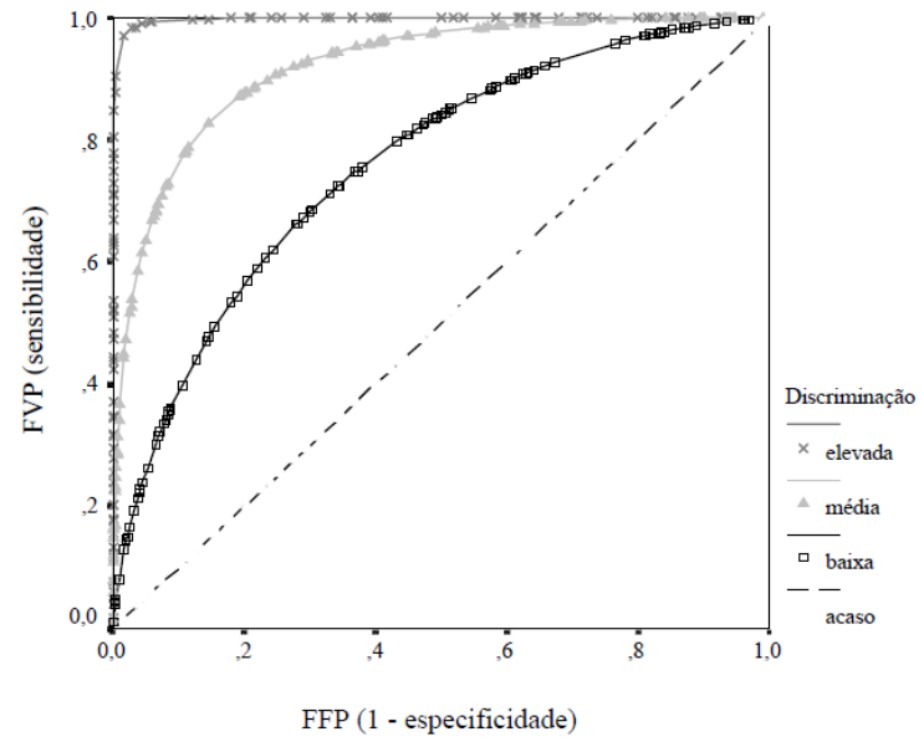

Figura 2. 4: Gráfico da Curva de ROC

Fonte: SANTA CATARINA (2009), adaptado de BRAGA (2000)

A área sob a curva ROC (AUC – *Area Under the Curve*) é a medida utilizada para sumarizar a qualidade da curva. Quanto mais a AUC aproximar-se de 1 melhor o desempenho. Este método é bastante utilizado porque é uma medida global de desempenho

independente de limites de corte, geralmente empregados na construção da matriz de confusão (DELEO, 1993).

2.2 Ajustes dos SDMs

Para ajustar SDMs, existem aplicações livres como o programa *openModeller.* Este aplicativo é uma ferramenta livre, de código aberto, para modelagem de distribuição espacial, desenvolvida pelo Centro de Referência em Informação Ambiental – CRIA (*OPENMODELLER*, 2008).

No programa *openModeller* estão implementados diversos algoritmos para a geração de SDMs, um exemplo desses algoritmos é o algoritmo GARP - *Genetic Algorithm for Rule-set Prediction*, que é utilizado mais precisamente na predição da distribuição potencial de espécies. Este algoritmo não considera os relacionamentos espaciais e, consequentemente, a dependência espacial; ele ajusta modelos e realiza predições observando apenas os valores pontuais das amostras (SANTA CATARINA, 2009).

O algoritmo GARP, é implementado no *openModeller* em duas variantes, o Single-Run e o *best-subset.* O algoritmo GARP *best-subset* ajusta vários modelos GARP Single-Run. Ao final do processo, um número pré-determinado de melhores modelos Single-Run são selecionados para compor o mapa de distribuição potencial final (BORGUESAN, 2013).

2.2.1 Best-Subset

Como o nome do método já diz ele tem como objetivo selecionar o melhor subconjunto de resultados baseado em alguns parâmetros pré-definidos (BORGUESAN, 2013).

O método *best-subset*, consiste em avaliar “n” modelos ajustados, seguindo os cinco passos descritos a seguir.

1. Ordenar, de modo crescente, os “n” modelos pela sua taxa de erros de omissão;

2. Selecionar apenas aqueles que apresentam taxas de erros de omissão inferiores ao limite máximo admitido nos parâmetros, evitando que modelos com altas taxas de erro grave sejam escolhidos;

3. Os modelos selecionados no passo anterior são ordenados pela sua taxa de erros de comissão;

4. Selecionar os “m” modelos mais próximos do valor mediano das taxas de erro de omissão, evitando modelos com superajuste, que é quando o modelo prediz precisamente cada ponto de presença, e modelos com superpredição, que é quando o modelo prediz área de presença em quase todo o mapa;

5. Construção de um mapa médio baseado nos “m” melhores modelos gerados. Este mapa médio corresponde ao SDM gerado através do método *best-subset*.

Neste método ajusta-se um número elevado de modelos, geralmente mais de 100 e, a partir destes, selecionam-se os melhores modelos para então criar um SDM médio que melhor represente a distribuição da espécie em estudo, logo é particularmente útil fazer o uso de métodos de pré-análise de dados reduzindo a dimensão do espaço de características e consequentemente acelerando o processo de ajuste do modelo.

2.3 SAHGA SDM

O sistema SAHGA SDM é um sistema que emprega o algoritmo SAHGA para modelar a distribuição potencial de espécies. O diferencial deste sistema está na sua capacidade de construir SDMs que considerem os relacionamentos espaciais presentes nos dados de entrada, representando-os através de uma MPG (matriz de proximidade generalizada) (SANTA CATARINA, 2009).

A MPG, ou matriz de proximidade generalizada, é uma variação da matriz de proximidade. Os pesos são calculados a partir de relações espaciais no espaço absoluto como distância euclidiana e adjacência, ou com base em relações espaciais no espaço relativo, que levam em conta a conectividade de objetos em uma rede de transporte ou de comunicação, por exemplo (AGUIAR et al., 2003) (PEDROSA, 2003) (SANTA CATARINA, 2009).

Uma MPG é composta por um conjunto de objetos geoespaciais “O” que são representados por células regulares ou polígonos, de acordo com a representação utilizada; um grafo “G” que é constituído por um conjunto de nós e arcos, onde cada nó representa um objeto e os arcos representam os relacionamentos de vizinhança entre dois nós; e uma matriz de

proximidade “V” que indica o quão próximo dois objetos “O” estão, geralmente representada em termos de adjacência ou distância euclidiana (SANTA CATARINA, 2009).

Os dados de entrada para o SAHGA SDM são: pontos amostrais de presença ou ausência da espécie, com seus relacionamentos espaciais (MPG), e o conjunto de *layers* geográficos que representam as variáveis ambientais que podem delimitar a sobrevivência da espécie (SANTA CATARINA, 2009).

Capítulo 3

Metodologia

Neste capítulo serão descritos os métodos de pré-análise que foram implementados e avaliados no processo de construção de SDMs. O método Jackknife foi utilizado para que, de modo rápido, porém eficiente, para cada ajuste do modelo foi feito o corte de uma variável geográfica assim resultado o quanto esta variável é importante para o modelo, logo podendo realizar redução de variáveis ambientais de entrada realizando a exclusão das variáveis menos importantes sem afetar a precisão do modelo. Visando o mesmo objetivo, de reduzir a quantidade de variáveis ambientais, o método de Correlação Linear foi implantado, esperando que hajam correlações significativas entre duas variáveis possibilitando a retirada de uma sem prejudicar o ajuste do mesmo. De maneira conjunta, porém contrária, a metodologia do Chi-Quadrado informa o quão dispersas são duas variáveis observadas, dos valores esperados, sem poder excluí-las do estudo.

Por fim, *FrameWork Qt Creator* foi usado para remodelar a interface do sistema SAHGA SDM, assim simplificando sua utilização.

3.1 Método de pré-análise

O processo para gerar um modelo de distribuição de espécies é complexo, pois requer manipular uma grande quantidade de dados de entrada, que muitas vezes é fornecido para o algoritmo de modelagem. Este volume de dados está relacionado principalmente com as variáveis ambientais georreferenciadas. Além da quantidade de variáveis compondo conjuntos volumosos, os arquivos através dos quais essas variáveis são disponibilizadas geralmente são grandes. Problemas com o tamanho dos arquivos podem ser contornados recortando a região de interesse e utilizando apenas esse recorte como arquivo de entrada, cujo tamanho pode ser consideravelmente menor (PINAYA, 2013).

Essa característica, definida pelo volume do conjunto de dados de entrada, está relacionada com a dimensão do espaço de características. Quanto maior o espaço de características, maior sua complexidade e, consequentemente, o tempo de processamento dos algoritmos na busca por associações entre os dados tende a ser maior. Além disso, os algoritmos induzirão modelos mais complexos (PINAYA, 2013).

Uma solução para minimizar esses problemas é a utilização de alguma técnica que auxilie o usuário na tarefa de selecionar um subconjunto de variáveis ambientais dentre todo o conjunto disponível.

Descreve-se a seguir, os métodos de pré-análise para aprimoramento do resultado esperado no sistema SAHGA SDM.

3.1.1 Jackknife

O teste Jackknife mede a importância das variáveis, estimando o ganho quando a variável é aplicada isolada e a perda quando é omitida. Assim, as variáveis preditoras podem ser classificadas de acordo com sua contribuição relativa para a elaboração do modelo. Trata-se de uma técnica de amostragem, pois utiliza subamostras construídas a partir da amostra original de dados, utilizadas para calcular as estimativas (RODRIGUES, 2012).

O Jackknife pode ser descrito resumidamente como segue:

Seja θ o parâmetro de interesse a ser estimado. Seleciona-se uma amostra original de dados *X* de tamanho *n*:

$$x_{(i)} = \{x_1, x_2, x_3, \dots, x_n\}. \quad (3.1)$$

Produz-se *n* amostras a partir da amostra original *X*, eliminando-se o i-ésimo exemplo (*i* = 1, ..., *n*) em cada nova amostra:

$$x = \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}. \quad (3.2)$$

Simplificando, o método de pré-análise Jackknife ajusta “n” modelos, suprimindo uma das variáveis ambientais ou climáticas a cada execução. Posteriormente os modelos são avaliados para aferir a importância de cada variável suprimida sobre a qualidade do modelo ajustado.

3.1.2 Correlação Linear

O estudo da correlação tem por objetivo medir o grau de relação existente entre duas ou mais variáveis aleatórias, isto é, saber se as alterações sofridas por uma das variáveis são acompanhadas de alterações nas outras. Uma vez caracterizada esta relação procura-se descreve-la de forma matemática, através de uma função. A estimação dos parâmetros dessa função matemática é o objeto da regressão linear. Os pares de valores de duas variáveis serão colocados num diagrama cartesiano chamado de “diagrama de dispersão”, em que, sua vantagem baseia-se na simples observação do gráfico que proporciona uma ideia bastante clara de como as duas variáveis se relacionam. Uma medida do grau e do sinal é dada pela covariância entre as duas variáveis aleatórias X e Y que é uma medida numérica de associação linear existente entre elas, e definida por:

$$Cov(X,Y) = \frac{1}{n}.[\Sigma x.y - \frac{\Sigma x.\Sigma y}{n}] \tag{3.3}$$

É mais conveniente usar para medida de correlação, o coeficiente de correlação linear de Pearson, como estimador de ρxy, definido por:

$$\rho xy = \frac{Cov(x,y)}{\sqrt{\sigma_x.\sigma_y}} \tag{3.4}$$

O coeficiente de correlação linear é um número puro que varia de -1 a +1 e sua interpretação depende do valor numérico e do sinal, como segue:

xy = -1 ⇨ Correlação perfeita negativa.

-1 < xy < 0 ⇨ Correlação negativa.

xy = 0	⇨	Correlação nula.
0 < xy < 1	⇨	Correlação positiva.
xy = 1	⇨	Correlação perfeita positiva.
0,2 < xy < 0,4	⇨	Correlação fraca*
0,4 < xy < 0,7	⇨	Correlação moderada*
0,7 < xy < 0,9	⇨	Correlação forte*

*possui o mesmo significado para os casos positivos e negativos (SPIEGEL, 1974).

Assim, por exemplo, pode-se medir se a relação entre o número de filhos de uma família e sua renda é forte, fraca ou nula.

Neste trabalho a Correlação Linear foi utilizada para medir o grau de associação entre pares de variáveis ambientais ou climáticas amostradas em pontos de presença da espécie. Ou seja, para cada ponto de presença constroem-se tuplas cujos valores correspondem às variáveis ambientais ou climáticas na coordenada geográfica correspondente a um ponto de presença da espécie. Pares de variáveis cujos valores estão nas tuplas tem seu coeficiente de correlação linear de Pearson calculado. Se este valor indicar uma correlação forte uma das variáveis pode ser excluída do processo de modelagem.

3.1.3 Chi-Quadrado

Segundo Toledo e Ovalle (1985), Chi-Quadrado é um teste de hipóteses que se destina a avaliar a dispersão para duas variáveis nominais, medindo a associação existente entre variáveis qualitativas. É utilizado para verificar a frequência com que um determinado acontecimento observado em uma amostra se desvia significamente ou não da frequência com que ele é esperado. Sendo que o pesquisador trabalha com duas hipóteses:

• Hipótese nula: Não há associação entre os grupos, ou seja, as variáveis são independentes.

• Hipótese alternativa: Há associação entre os grupos, ou seja, as variáveis são dependentes.

De maneira estatística X^2 (leia-se: Qui-Quadrado), é expressa por:

$$\chi^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \quad (3.5)$$

Onde *e* é a frequência esperada e *o* é a frequência observada.

Após aplicar o método do Chi-Quadrado, obtendo o valor de Chi, é utilizada a tabela de distribuição de Chi-Quadrado a partir do grau de liberdade (equação 3.4) e nível de confiança igual a 95%. Compara-se o valor Chi calculado com o valor tabelado; se o valor Chi calculado for maior que o valor obtido na tabela, as variáveis são dependentes caso contrário as variáveis são independentes.

$$GL = (C - 1) * (L - 1) \quad (3.6)$$

Onde C é o número de colunas da tabela e L é o número de linhas da tabela.

O teste Chi-Quadrado é aplicado em pares de variáveis climáticas ou ambientais amostradas nos pontos de presença. A partir dos resultados do Chi-Quadrado, pode-se fazer a exclusão de variáveis dependentes.

Para melhor compreensão do método Qui-Quadrado, um exemplo de aplicação do teste é apresentado no Apêndice A.

3.2 Qt Creator

Para implementar os métodos de pré-análises no sistema SAHGA SDM, foi utilizado o kit de desenvolvimento de sistemas Qt Creator, principalmente porque o sistema está implementado nesta plataforma em linguagem C++.

Qt Creator é um IDE voltado para programação em Qt (biblioteca gráfica) na linguagem C++. Com ele, consegue-se programar de forma mais rápida e fácil, pois o programa traz alguns recursos, como o auto-completar, que facilitam muito o trabalho do programador.

Um exemplo da abrangência que esta biblioteca possui é o projeto KDE, que vem evoluindo de forma rápida e eficiente. A principal vantagem mostrada pela Qt é a sua portabilidade, ou seja, a possibilidade de rodar um mesmo programa nos diversos Sistemas Operacionais presentes no mercado como: Linux, Unix, MacOs e Windows (MARTINS, 2014).

Capítulo 4

Estudo de Caso

Para os parâmetros de entrada foram usados os dados da espécie *Thalurania furcata boliviana* (Figura 4.1), com 65 pontos de presença e 8 *layers* geográficos correspondentes às variáveis: precipitação acumulada no trimestre mais úmido, precipitação acumulada no trimestre mais quente, precipitação anual, temperatura média anual, temperatura média no trimestre mais frio, temperatura média no trimestre mais seco, temperatura média no trimestre mais quente e temperatura média no trimestre mais úmido. Para geração de pontos de pseudoausência utilizou-se o algoritmo BIOCLIM, implementado no sistema; foram gerados 65 pontos de pseudoausência, criando um conjunto amostral balanceado.

Figura 4. 1: *Thalurania furcata boliviana*

Fonte: BARTLEY (2009)

Na geração da MPG utilizou-se o raio de 60 km, ou seja, se dois pontos, presença/ausência, estão distantes em até metade do raio, estes pontos estão relacionados com peso 1; se os dois pontos estarão distantes entre 30 km e 60 km o peso destes pontos é de 0.5; para distâncias maiores que 60 km o peso do relacionamento é nulo.

Nos parâmetros, para o algoritmo genético, foi utilizado para realizar o ajuste do modelo o parâmetro Hard, por ser um conjunto de parâmetros genéticos que visa assegurar ao algoritmo uma qualidade da resposta com tempo de convergência aceitável (SANTA CATARINA, 2009). E para ajustar os modelos optou-se pelo tipo linear, por terem se mostrado suficientes para relacionar a ocorrência da espécie com as variáveis ambientais, para a espécie escolhida.

Para a função objetivo fez-se o uso do tipo Min-Both, que implementa a maximização do ajuste do modelo e também o acerto global do modelo (SANTA CATARINA, 2009) (FINAMORE, 2010).

4.1 Resultados

Os resultados foram divididos em quatro seções; a primeira com os resultados do algoritmo Jackknife, a segunda com os resultados do algoritmo Correlação Linear, a terceira os resultados do algoritmo Chi-Quadrado e a quarta e última seção apresenta a nova interface do sistema SAHGA SDM.

4.1.1 Jackknife

Utilizando o algoritmo de pré-análise Jackknife, que em cada execução suprime uma variável ambiental para se obter o quanto ela é importante para o modelo, foram coletados os resultados apresentados na tabela 4.1.

Tabela 4. 1: SDMs ajustados utilizando o algoritmo de pré-análise Jackknife.

Variáveis excluídas	**Acurácia (%)**
Precipitação acumulada no trimestre mais úmido	77.27
Precipitação acumulada no trimestre mais quente	85.00
Precipitação anual	77.27
Temperatura média anual	85.00
Temperatura média no trimestre mais frio	77.27
Temperatura média no trimestre mais seco	77.27
Temperatura média no trimestre mais quente	93.77
Temperatura média no trimestre mais úmido	84.21

Analisando a tabela 4.1, pode-se perceber o quanto cada variável climática (layer geográfico) é importante para o modelo. Por exemplo, a variável "Temperatura média no trimestre mais quente" é a menos importante para modelar a distribuição da espécie em estudo pois mesmo realizando a sua exclusão a acurácia continuou com um valor elevado; assim pode-se excluir esta variável. Na tabela 4.2 são apresentadas a acurácia e o tempo de execução do SAHGA SDM, com o algoritmo *best-subset*, para o modelo com todas as variáveis ambientais e para o modelo onde foi excluída a variável "Temperatura média no trimestre mais quente".

Tabela 4. 2: Comparação SDMs com exclusão de *layers* com o método Jackknife

Layers excluídas	**Acurácia (%)**	**Tempo (minutos)**
Nenhuma	92.91	9.61
Temperatura média no trimestre mais quente	91.98	8.86

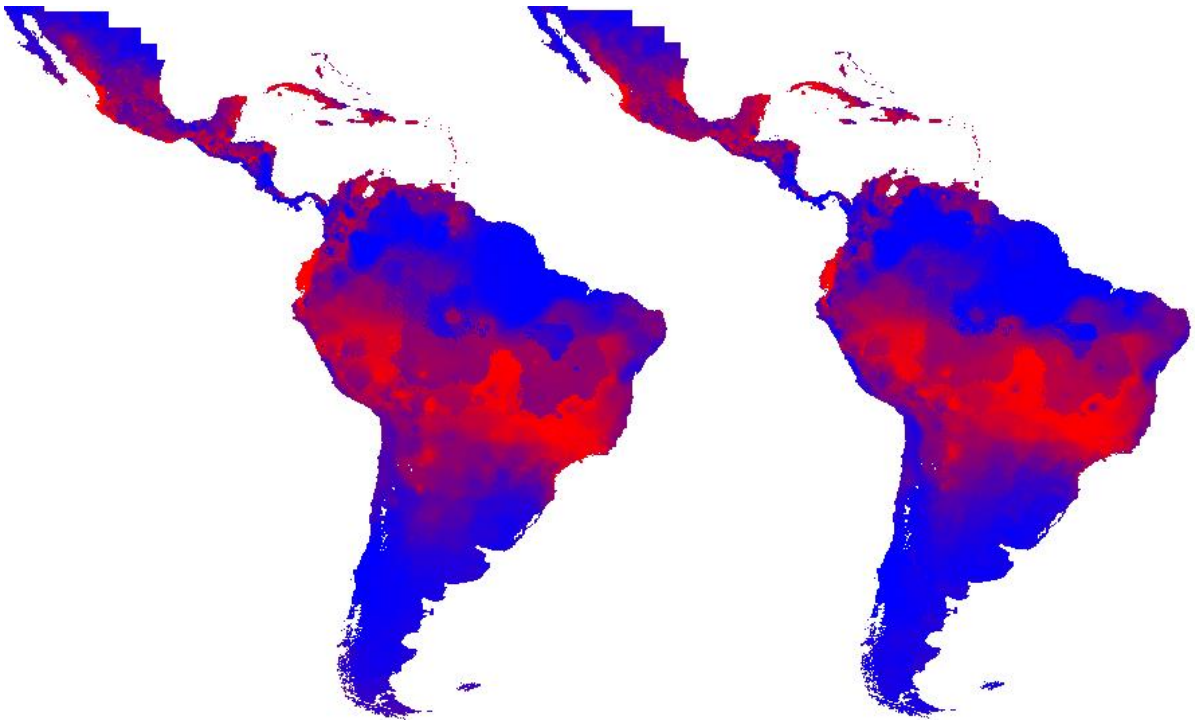

Figura 4. 2: Comparação do ajuste do modelo com o método Jackknife

Os dados apresentados na tabela 4.2 mostram que a exclusão de uma variável de entrada reduziu em quase 1 minuto o tempo de modelagem, sem reduzir significativamente a acurácia

do modelo ajustado. Os mapas de distribuição correspondentes aos modelos ajustados são apresentados na figura 4.2. A porção esquerda da imagem apresenta o modelo onde foi excluída uma variável ambiental; a porção direita da imagem mostra o modelo ajustado com todas as variáveis disponíveis e pode-se perceber que não houve mudança no mapa de distribuição de espécies.

4.1.2 Correlação Linear

Utilizando o algoritmo de pré-análise Correlação Linear, algoritmo que para cada par de variáveis ambientais mede o seu grau de correlação, foram coletados os resultados apresentados na tabela 4.3.

Tabela 4. 3: Correlação entre as variáveis geográficas

Layer	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**
L1	0.950012	0.957873	0.685312	0.716708	0.73371	0.716106	0.722615
L2		0.993163	0.889374	0.896466	0.899834	0.88698	0.889046
L3			0.878482	0.886236	0.889855	0.87388	0.8754
L4				0.999344	0.999215	0.999697	0.999687
L5					0.999772	0.998177	0.998167
L6						0.998104	0.998127
L7							0.999978

Onde:

- L1: Precipitação acumulada no trimestre mais úmido;
- L2: Precipitação acumulada no trimestre mais quente;
- L3: Precipitação anual;
- L4: Temperatura média anual;
- L5: Temperatura média no trimestre mais frio;
- L6: Temperatura média no trimestre mais seco;
- L7: Temperatura média no trimestre mais quente;
- L8: Temperatura média no trimestre mais úmido.

A partir da análise dos dados apresentados na tabela 4.3, pode-se perceber o grau de correlação entre as variáveis utilizadas no modelo. Analisando a tabela percebemos que existe uma correlação forte entre as variáveis L4, L5, L6, L7 e L8, assim podendo realizar a exclusão de 4 entre essas 5 variáveis sem afetar o modelo. A tabela 4.4 apresenta os resultados obtidos com a exclusão dessas de 4 variáveis geográficas e utilizando algoritmo *best-subset* para o ajuste do modelo pois o mesmo fornece resultados mais confiáveis.

Tabela 4. 4: Comparação SDMs com exclusão de layers com o método Correlação Linear

Layers excluídas	**Acurácia (%)**	**Tempo (minutos)**
Nenhuma	92.91	9.61
L5, L6, L7, L8	93.27	6.46

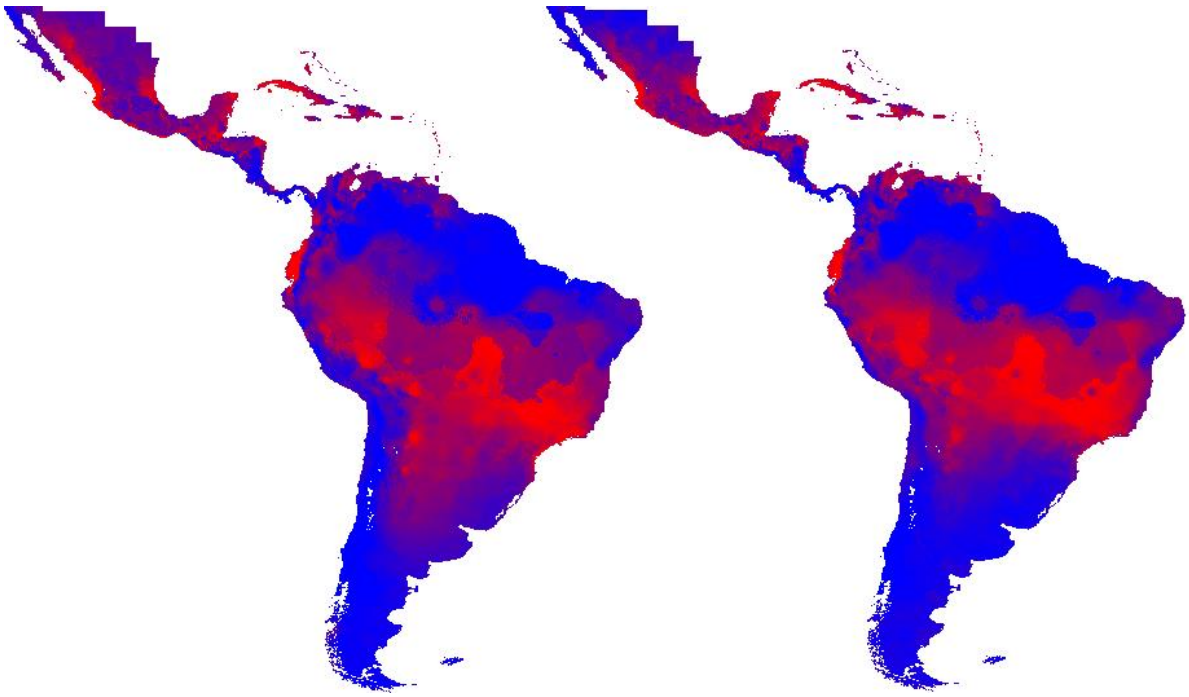

Figura 4. 3: Comparação do ajuste do modelo com o método Correlação Linear

Os dados apresentados na tabela 4.4 mostram que a exclusão de quatro variáveis de entrada reduziu em aproximadamente 3 minutos o tempo de modelagem, sem alterar significativamente a acurácia do modelo ajustado. Os mapas de distribuição correspondentes aos modelos ajustados são apresentados na figura 4.3. A porção esquerda da imagem apresenta o modelo onde foi excluída as variáveis ambientais; a porção direita da imagem

mostra o modelo ajustado com todas as variáveis disponíveis e pode-se perceber que houve uma pequena alteração no mapa de distribuição de espécie mas nada que afete o modelo ajustado.

4.1.3 Chi-Quadrado

Utilizando o algoritmo de pré-análise Chi-Quadrado, algoritmo que testa a independência condicional entre as variáveis ambientais, foram coletados os resultados apresentados na tabela 4.5.

Tabela 4. 5: Resultado do método Chi-Quadrado

Layer	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**
L1	Depend.	Depend.	Depend.	Depend.	Depend.	Depend.	Depend.
L2		Depend.	Depend.	Depend.	Depend.	Depend.	Depend.
L3			Depend.	Depend.	Depend.	Depend.	Depend.
L4				Depend.	Depend.	Depend.	Depend.
L5					Depend.	Depend.	Depend.
L6						Depend.	Depend.
L7							Depend.

Onde:

- L1: Precipitação acumulada no trimestre mais úmido;
- L2: Precipitação acumulada no trimestre mais quente;
- L3: Precipitação anual;
- L4: Temperatura média anual;
- L5: Temperatura média no trimestre mais frio;
- L6: Temperatura média no trimestre mais seco;
- L7: Temperatura média no trimestre mais quente;
- L8: Temperatura média no trimestre mais úmido.

As variáveis são dependentes, ou seja, a distribuição de pares (Vi; Vj) não acontecem por acaso portanto estas variáveis não devem ser suprimidas do modelo.

Entretanto, devido ao baixo número de amostras, os resultados obtidos com o Chi-Quadrado, não são confiáveis. Segundo Fisher (1934), o número de observações em cada tupla deveria ser maior ou igual a 5.

Neste estudo o valor máximo encontrado nas tuplas foi 2 e muitas com valor 0.

4.1.4 Interface SAHGA SDM

Com o uso do *framework* Qt Creator, foi concluída a remodelagem do sistema SAHGA SDM. Na figura 4.4 é apresentado a tela principal do sistema em sua interface antiga, já na figura 4.5 é apresentado a tela principal com a nova interface do sistema. Nota-se que em sua interface antiga, o resultado da modelagem não era exibido em tela, pois era salvo em um arquivo HTML e visualizado através de um navegador; já na nova interface o resultado da modelagem é apresentado em tela, facilitando a visualização da modelagem.

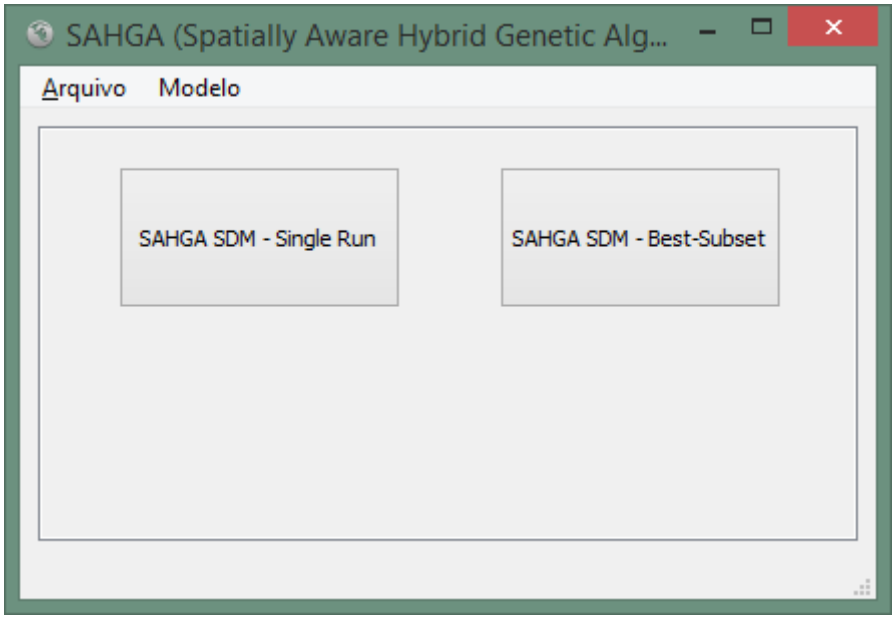

Figura 4. 4: Tela principal da interface antiga do SAHGA SDM

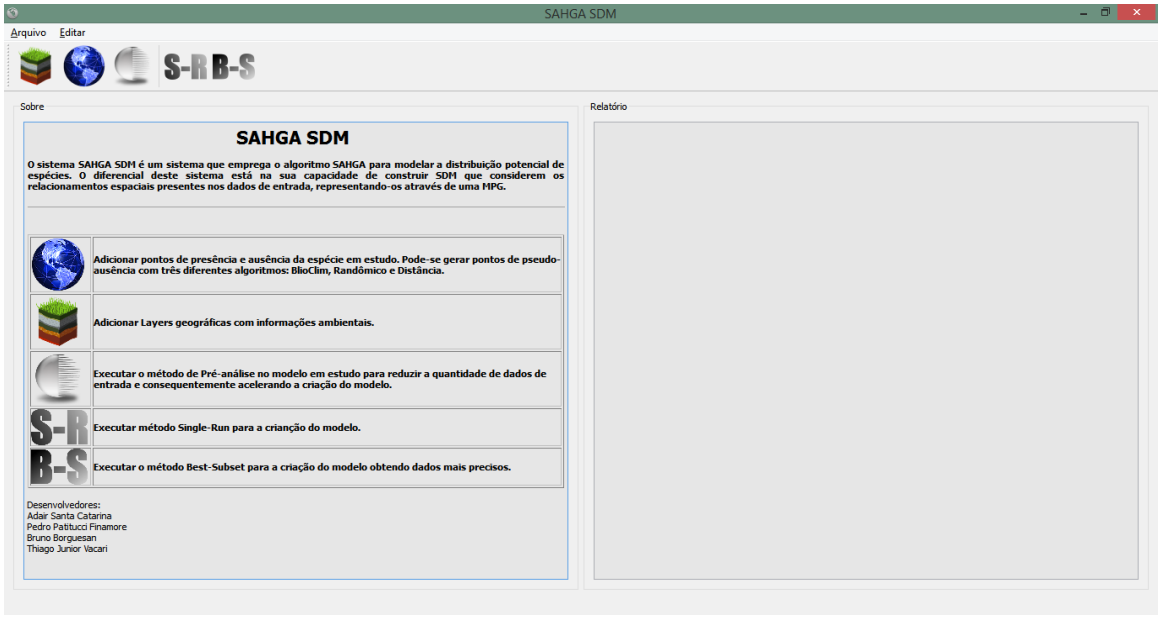

Figura 4. 5: Tela principal do SAHGA SDM

Em seguira será apresentado as demais telas do sistema SAHGA SDM com sua interface remodelada.

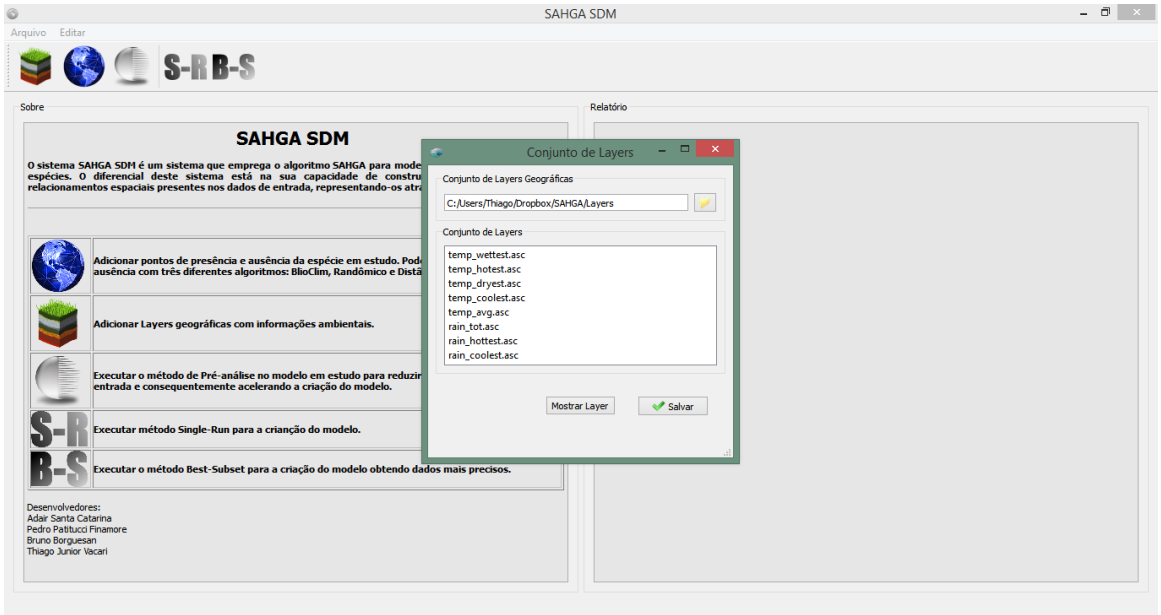

Figura 4. 6: Janela para carregar variáveis geográficas

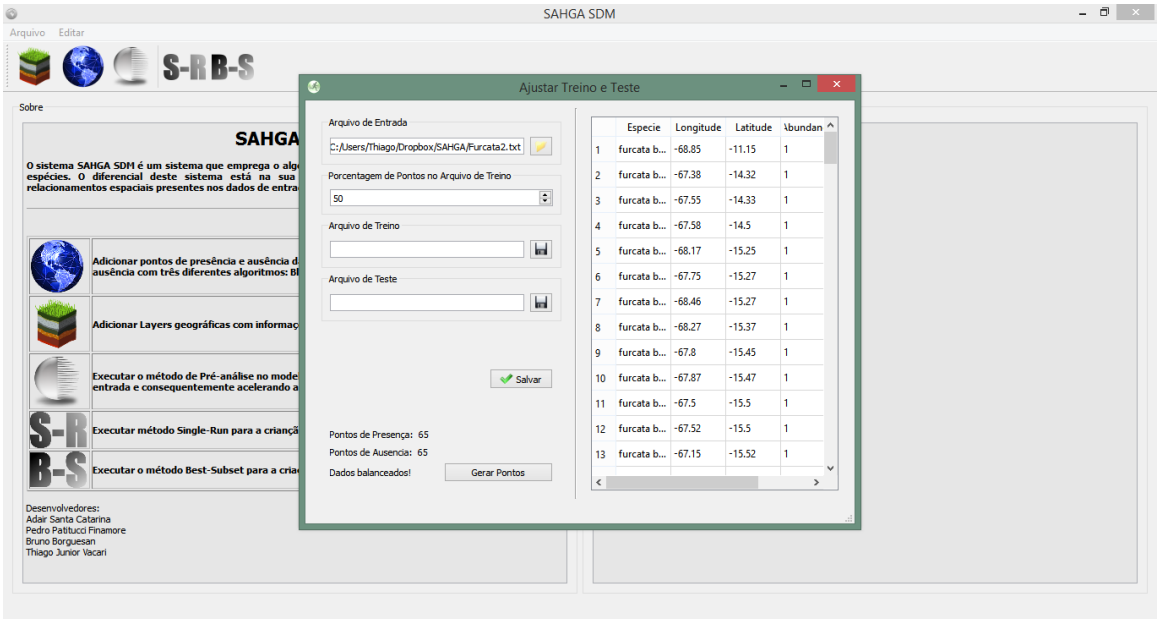

Figura 4. 7: Janela para carregar pontos de presencia/ausência da espécie

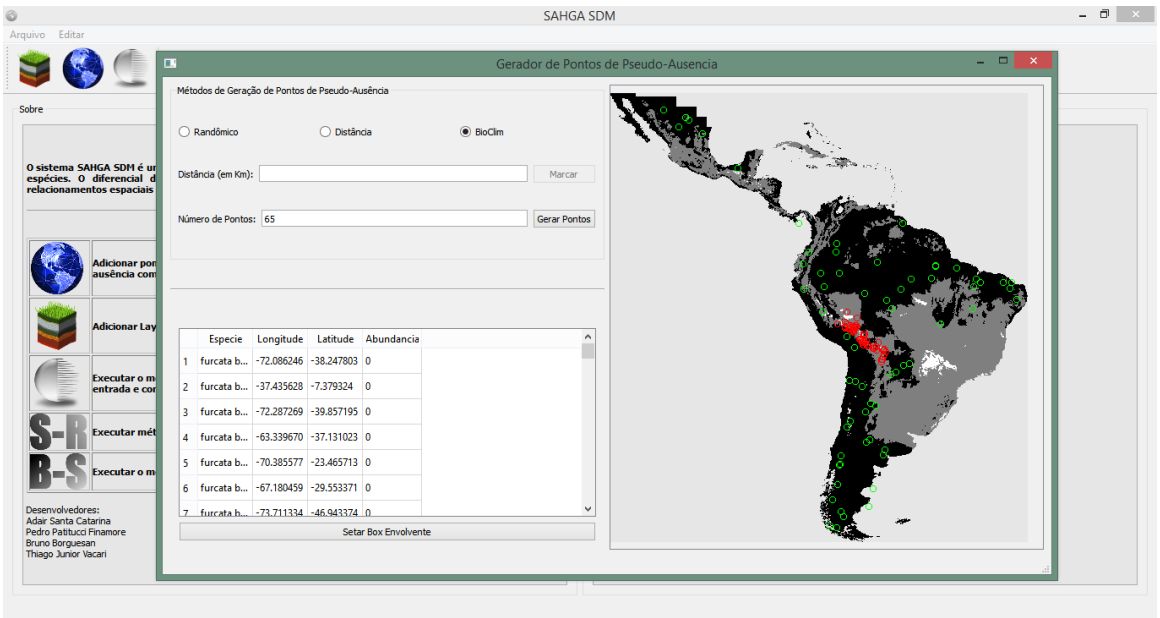

Figura 4. 8: Janela para a geração de pontos de pseudoausência

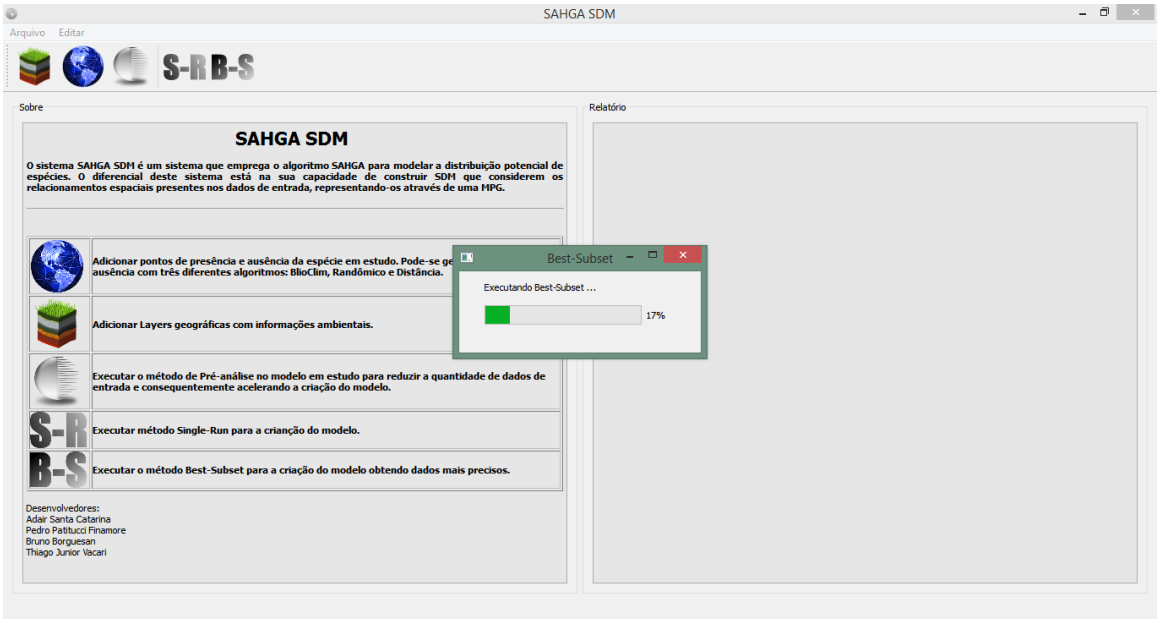

Figura 4. 9: Janela de execução do método *best-subset*

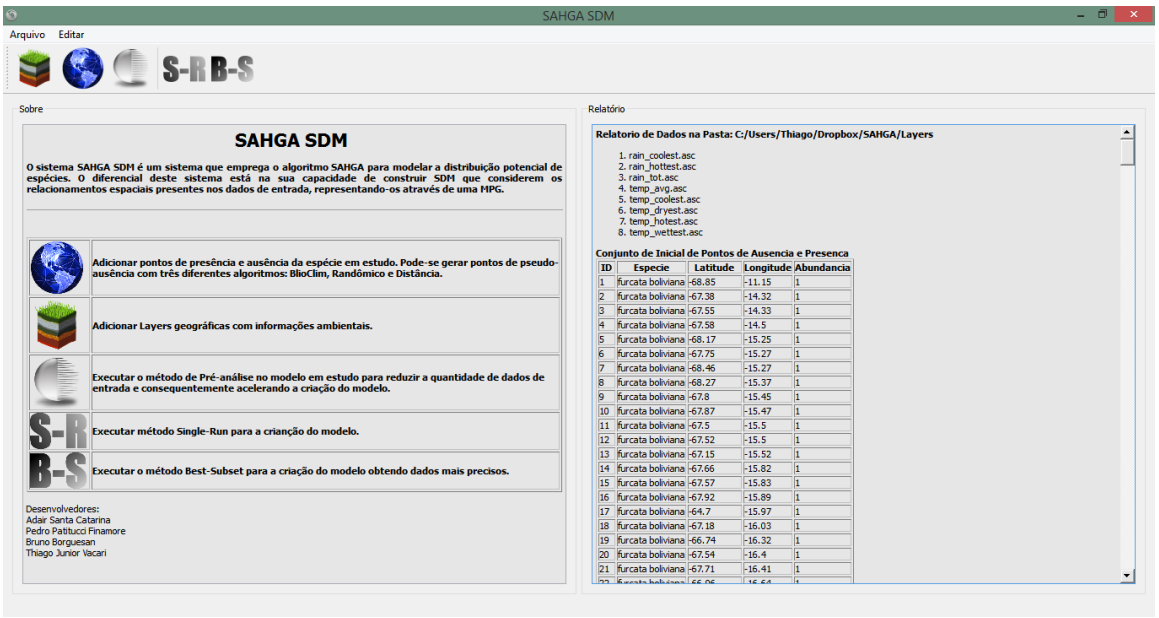

Figura 4. 10: Janela principal com os resultados da modelagem de distribuição de espécies com os pontos de presença e as variáveis ambientais usadas na modelagem

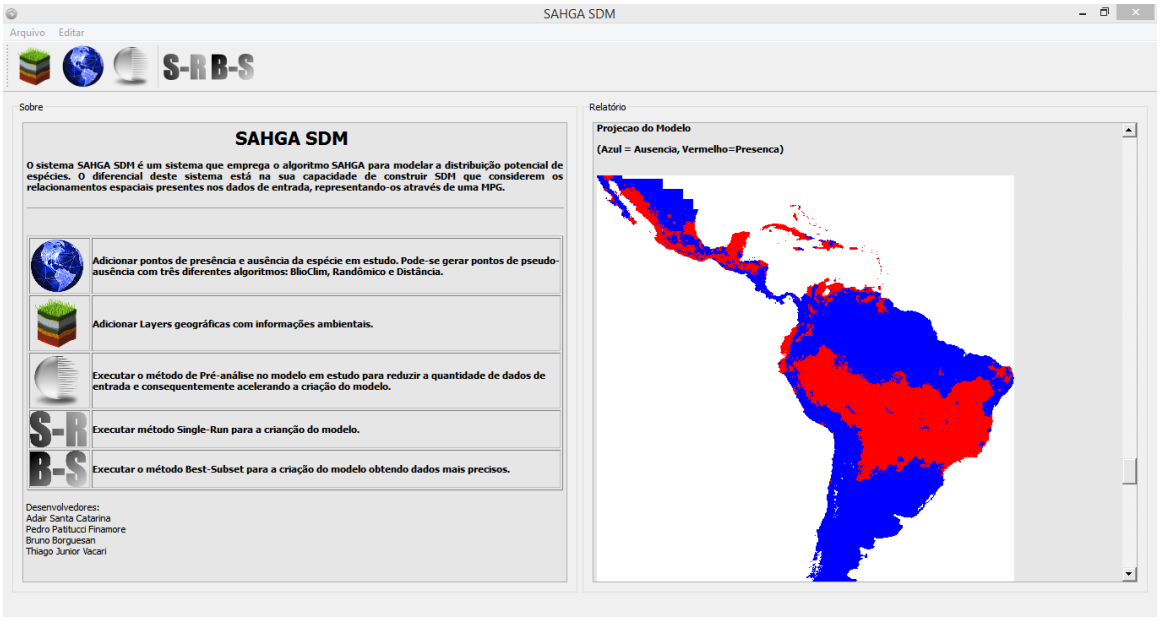

Figura 4. 11: Janela principal com os resultados da modelagem de distribuição de espécies com a projeção do modelo ajustado

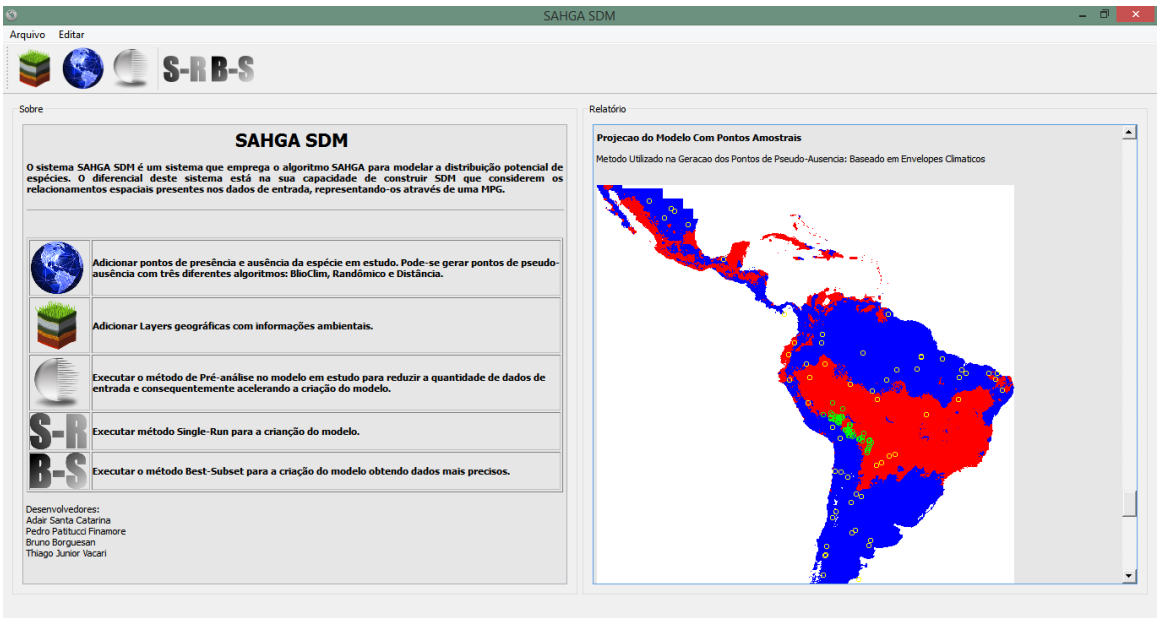

Figura 4. 12: Janela principal com os resultados da modelagem de distribuição de espécies com a projeção do modelo ajustado com pontos amostrais

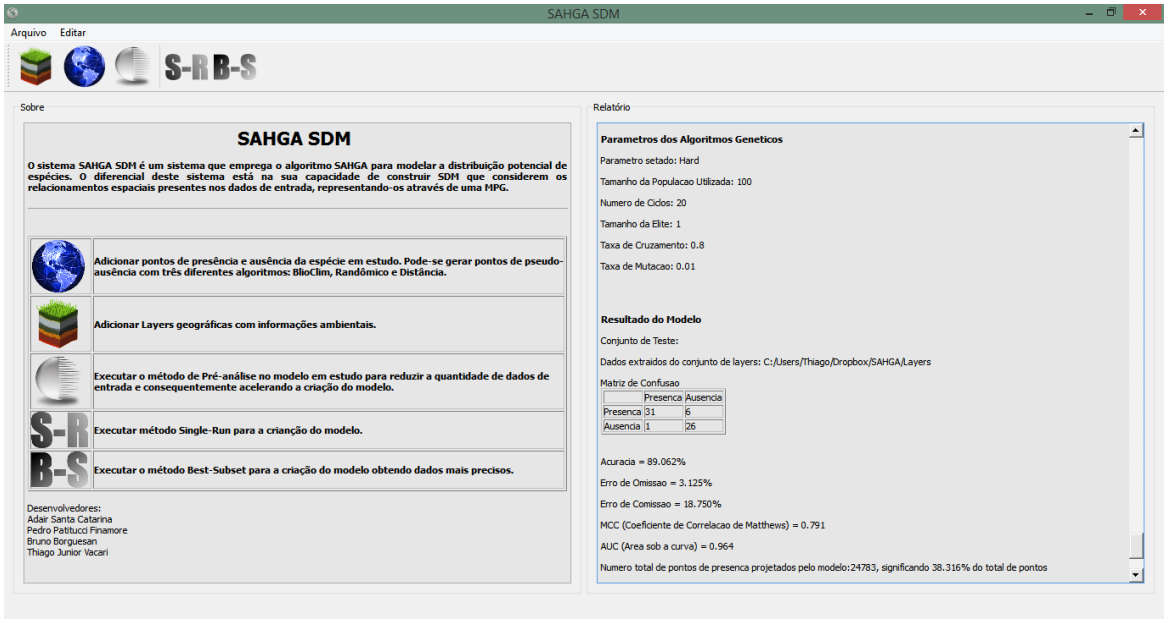

Figura 4. 13: Janela principal com os resultados da modelagem de distribuição de espécies com a matriz de confusão e algumas medidas de avaliação do modelo

Para as janelas de carregar *layers*, carregar pontos de presencia/ausência e de geração de pontos de pseudoausência, figuras 4.6, 4.7 e 4.8 respectivamente, foram reusadas da interface antiga do sistema, fazendo apenas a importação para a nova versão do *framework Qt Creator,* para isso tiveram que ser alterados algumas importações de bibliotecas da antiga versão do sistema SAHGA SDM para serem suportadas na nova versão do *framework Qt Creator*.

Em sua nova interface, o sistema SAHGA SDM exibe os resultados das modelagens direto em sua tela principal, apresentado nas figuras 4.10, 4.11, 4,12 e 4.13, diferente da antiga versão que salvava os resultados em um arquivo e então o usuário teria o trabalho de abri-lo.

E por fim, como a execução do método *best-subset* é muito custoso, podendo chegar a horas dependendo o número de modelos e de variáveis de entrada, foi implantado uma janela com um *progress bar* com o uso de *multithreads* para fazer a atualização da tela apresentado na figura 4.9.

Capítulo 5

Conclusões

Neste trabalho foram implementados três diferentes métodos de pré-análise de dados: Jackknife, Correlação Linear e Chi-Quadrado, integrando-os no sistema de modelagem de distribuição de espécies SAHGA SDM.

O algoritmo de pré-análise Jackknife apresentou resultados animadores. Para os dados da espécie *Thalurania furcata boliviana* percebeu-se que a retirada de uma das oito variáveis utilizadas reduziu em 1 minuto o tempo de ajuste do modelo sem prejuízos em sua acurácia.

O algoritmo de pré-análise Correlação Linear identificou que 50% das variáveis apresentavam correlação acima de 99,9% o que permitiu ajustar o modelo com apenas quatro variáveis. O tempo de modelagem foi reduzido em aproximadamente 3 minutos com um pequeno incremento na acurácia (de 92,91% para 93,27%).

Por fim o método de pré-análise Chi-Quadrado não se mostrou adequado devido ao pequeno número de pontos de presença, o que inviabiliza as conclusões do teste.

Conclui-se, portanto que os métodos de pré-análise Jackknife e Correlação Linear são adequados para serem integrados ao sistema SAHGA SDM pois foram os métodos de pré-análise que obtiveram os melhores resultados.

5.1 Trabalhos Futuros

Ao concluir esse trabalho percebeu-se que o sistema SAHGA SDM pode evoluir e ter algumas funcionalidades adicionadas. São desafios para trabalhos futuros:

- Adicionar suporte ao número maior de tipos de arquivos de entrada, pois atualmente o sistema faz o uso de somente variáveis de entrada arquivos do tipo asc.
- Estruturar e comentar o código fonte do sistema SAHGA SDM para facilitar sua manutenibilidade.

Apêndice A

Exemplo de aplicação do método Qui-Quadrado

O exemplo a seguir foi retirado do material de Conti (2011).

Os resultados abaixo, apresentados na tabela A.1, provêm de um teste sorológico aplicado a indivíduos pertencentes a 3 amostras compostas por indivíduos de provenientes de diferentes faixas etárias (crianças, adolescentes e adultos). Pôr à prova a hipótese de que a proporção de indivíduos com reação positiva não difere significativamente nas 3 amostras contra a hipótese de que isso não é verdadeiro.

Tabela A. 1: Tabela de amostras do dado

Amostra	Reação +	Reação -	Total
Crianças	25	45	70
Jovens	15	25	40
Adultos	10	30	40
Total	50	100	150

Para calcular os esperados multiplica-se os totais parciais relativos a cada casela e divide-se pelo total geral (N).

Por exemplo, na casela crianças + = 50 x 70 / 150 = 23,3333

Depois calcula-se os valores de Qui-Quadrado parciais.

Por exemplo, na casela crianças + = (o-e)2 /e = [(25 - 23,3333)2 / 23,3333)] = 0,1190.

Depois, calcula-se a parcela de X^2 referente a cada casela.

Ao final, soma-se as parcelas e obtém-se o X^2, simplificado na tabela A.2.

Tabela A. 2: Valores de Qui-Quadrado

Amostra	**Reação +**	**Reação -**	**Total**
Crianças	25	45	70
Esp	23,3333	46,6667	
$(o-e)^2/e$	0,1190	0,0595	
Jovens	15	25	40
Esp	13,3333	26,6667	
$(o-e)^2/e$	0,2083	0,1042	
Adultos	10	30	40
Esp	13,3333	26,6667	
$(o-e)^2/e$	0,8333	0,4167	
Total	50	100	150

X^2 = 0,1190 + 0,0595 + 0,2083 + 0,1042 + 0,8333 + 0,4167. Portanto, X^2 = 1,7410.

O número de grau de liberdade em tabelas é assim calculado:

Grau de Liberdade = (número de linhas -1) x (número de colunas -1). Portanto: Grau de Liberdade = (2 - 1) x (3 - 1) = 2.

Depois, consulta-se a tabela de Qui-Quadrado e verifica-se que X^2 tabelado = 5,991.

Como o valor de X^2 calculado é menor que o valor de X^2 tabelado conclui-se que os desvios não são significativos.

Portanto, os indivíduos pertencentes às 3 amostras (crianças, adolescentes e adultos) reagem do mesmo modo ao teste sorológico, não havendo influência das diferentes faixas etárias sobre o resultado do teste. Assim sendo, o resultado sorológico independe dos grupos etários.

Referências Bibliográficas

AGUIAR, A. P. D. et al. Modelling Spatial Relations by Generalized Proximity Matrices. In: *GeoInfo*. 2003.

BARTLEY, G. *Birds of Bolivia. Nature Photogragraphy*, 2009. Disponivel em: <http://www.glennbartley.com/naturephotography/birds/EMERALDBELLIED%20WOODNYMPH.html>. Acesso em: 4 Outubro 2014.

BORGUESAN, B. *O Método Best-Subset Incorporado ao Sistema SAHGA SDM*. Monografia (Graduação em Ciência da Computação) – Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel, PR, 2013.

CASSEMIRO, F. A. S., GOUVEIA, S. F. & DINIZ-FILHO, J. A. F. Distribuição de Rhinella granulosa: integrando envelopes bioclimáticos e respostas ecofisiológicas *Revista da Biologia*, v. 8, p. 38–44. 2012.

CONTI, F. *Qui Quadrado*. 2011. Disponível em: <http://www.cultura.ufpa.br/dicas/biome/bioqui.htm>. Acesso em: 19 nov. 2014.

CORREA, S. M. B B. *Probabilidade e Estatística*. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003.

DELEO, J. M. Receiver operating haracteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: *Proceedings of the Second International symposium on Uncertainty Modelling and Analysis*, Los Alamitos, California: IEEE Computer Society Press, p. 318–325. 1993

FINAMORE, P. P. *Avaliação de Três Métodos para Geração de Pontos de PseudoAusência Sobre a Qualidade dos Modelos de Distribuição de Espécies Ajustados pelo Sistema SAHGA SDM*. Monografia (Graduação em Ciência da Computação) – Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel, PR, 2010.

Fisher, R.A. *Statistical Methods for Research Workers*. 5th Edition, Edinburgh: Oliver and Boyd, 1934.

GUISAN, A.; THUILLER, W. *Predicting Species Distribution: Offering more than Simple Habitat Models*. Ecology Letters, v. 8, n. 9, p. 993-1009, Jun. 2005.

GUISAN, A.; ZIMMERMANN, N. E. Predictive habitat distribution models in ecology. *Ecological Modelling*, v. 135, n. 2-3, p. 147-186, Dezembro 2000.

IWASHITA, F. *Sensibilidade de Modelos de Distribuição de Espécies a Erros de Posicio-Namento de Dados de Coleta*. 2007. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos.

LIRA, S. A. *Análise de Correlação: Abordagem Teórica e de Construção dos Coeficientes com Aplicações*. Dissertação (Mestrado em Ciências) – Universidade Federal do Paraná (UFPR), Curitiba, PR, 2004.

MARTINS, Elaine. *Qt Creator***:** Ótima IDE voltada para programação gráfica Qt em C++, Java e outras linguagens. 2014. Disponível em: < http://goo.gl/mSnMT>. Acesso em: 31 jul. 2014

NABOUT, João Carlos et al. *Distribuição Geográfica Potencial de Espécies Americanas do Carangueijo "violinista" (Uca spp.) (Crustacea, Decapoda) com Base em Modelagem de Nicho Ecológico*. 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0073-47212009000100013>. Acesso em: 31 jul. 2014

NIX, H. A. A Biogeographic Analysis of Australian Elapid Snakes. In: *LONGMORE, R.* (Ed.) Atlas of elapid snakes of Australia. Canberra: Australian government publishing service, v.7, 1986. p. 4–15. (Australian flora and fauna series)

CRIA, Centro de Referência em Informação Ambiental; POLI-USP, Escola Politécnica da USP; INPE, Instituto Nacional de Pesquisas Espaciais. *OpenModeller*. 2008. Disponível em: <http://openmodeller.sourceforge.net/>. Acesso em: 22/03/2014.

PEDROSA, B. M. *Ambiente Computacional para Modelagem Dinâmica*. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, 2003.

PINAYA, Jorge Luiz Diaz. *Processo de Pré-Análise para a Modelagem de Distribuição de Espécies.* 2013. 112 f. Dissertação (Mestrado) - Curso de Sistemas Digitais, Escola Politécnica da Universidade de São Paulo, São Paulo, 2013.

RODRIGUES, F. A. *Um Método de Referência para Análise de Desempenho Preditivo de Algoritmos de Modelagem de Distribuição de Espécies*. Tese (Doutorado em Ciências) – Universidade de São Paulo (USP), São Paulo, SP, 2012.

SANTA CATARINA, A. SAHGA – *Um Algoritmo Genético Híbrido com Representação Explícita de Relacionamentos Espaciais para Análise de Dados Geoespaciais*. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, 2009.

SIQUEIRA, M. F. *Uso de Modelagem de Nicho Fundamental na Avaliação do Padrão de Distribuição Geográfica de Espécies Vegetais*. Tese (Doutorado em Ciências de Engenharia Ambiental) - Universidade de São Paulo (USP), São Carlos, SP, 2005.

SPIEGEL, Murray R.. *Estatística*. São Paulo: Mcgraw-hill, 1974. 580 p

TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. *Estatística Básica*. 2. ed. São Paulo: Atlas, 1985. 459 p.