Abordagens de Modelagem Estatística e Aprendizado de Máquina para a Predição da Transmitância Atmosférica

José Rodrigues de Carvalho Neto[1],
Mauri Apacerido de Oliveira[1], Caio Augusto de Melo Arruda Silvestre[1,2]
[1]Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos/SP – Brasil
[2]Instituto de Aplicações Operacionais (IAOP), São José dos Campos/SP – Brasil

***Resumo* – Diante da crescente complexidade dos equipamentos militares, especialmente sistemas *Infra-Red Search and Track* (IRST), a FAB envidou esforços para desenvolver ferramentas para estimar o alcance desse tipo de sistema. Para tal, a transmitância atmosférica, inicialmente calculada pelo software MODTRAN, é essencial. Porém, o MODTRAN tem alta demanda computacional, comprometendo a agilidade das decisões. Assim, buscou-se criar um modelo matemático que previsse a transmitância de maneira mais rápida e econômica, utilizando dados do MODTRAN, especificamente na banda larga do espectro infravermelho. Inicialmente, a Regressão Linear Múltipla (RLM) foi testada, mas apresentou limitações. A regressão beta, apesar de limitar os valores entre 0 e 1, mostrou predição insatisfatória. O modelo Random Forest de aprendizagem de máquina apresentou excelentes resultados numa base de dados específica, porém, numa base com valores ligeiramente diferentes, a precisão caiu, revelando suas limitações.**

Palavras-Chave – Transmitância Atmosférica, Aprendizado de Máquina, Modelagem Estatística.

I. INTRODUÇÃO

No panorama de constantes avanços tecnológicos, sobretudo no domínio da guerra eletrônica, uma gama diversificada de armamentos vem emergindo e desempenhando papéis cruciais nos conflitos aéreos contemporâneos. Por cerca de cinquenta anos, o radar se firmou como o sensor de maior importância no campo de batalha, com uma evolução acelerada desde sua concepção.

Paralelamente, emergiram equipamentos e metodologias com o objetivo de impedir ou postergar a detecção de aeronaves por sistemas de radar, incluindo aparelhos de Jamming e a inovadora tecnologia *Stealth* [1]. Dentro desse contexto, um marco significativo foi a concepção dos sistemas *Infrared Search and Track* (IRST), que aproveitam a radiação infravermelha emitida por aeronaves para detectá-las e rastreá-las de forma passiva, ou seja, sem a necessidade de emissão eletromagnética, contrastando com os sistemas de radar tradicionais.

Essa capacidade se traduz em um valor operacional inestimável, permitindo interceptar ou confrontar outra aeronave sem emitir sinais eletromagnéticos e, consequentemente, sem ser detectada por receptores de alerta de radar ou RWR (*Radar Warning Receiver*) comumente integrados em aeronaves de combate. Considerando essa vantagem operacional significativa, tal tipo de sensor vem sendo incorporado em plataformas das forças aéreas de maior expressão mundialmente.

No Brasil, essa realidade está prestes a ser incorporada com a introdução das aeronaves Gripen E/F, que serão equipadas com o primeiro sensor IRST do país, o Skyward-G, desenvolvido pela empresa italiana Leonardo. Isso impulsiona a necessidade de elaborar a doutrina de emprego desse sistema.

Com essa perspectiva em mente, [2] desenvolveu um algoritmo capaz de estimar o alcance de detecção de sistemas IRST. Este algoritmo utiliza um modelo matemático fundamentado nos fenômenos físicos envolvidos. O modelo se vale de dados de transmitância atmosférica provenientes do MODTRAN e de valores de assinatura infravermelha de alvos obtidos a partir de simulações realizadas no SIMIS [3], o que confere ao modelo um alto grau de fidelidade e precisão.

Além disso nota-se que o algoritmo possui um alto custo computacional para a aquisição dos dados de transmitância atmosférica através do MODTRAN, o que poderia inviabilizar o uso do algoritmo de predição de alcance de detecção em algumas aplicações como o Ambiente de Simulação Aeroespacial (ASA) da FAB, cujas aplicações são exemplificadas em [4,5]. Dessa forma, manifesta-se a demanda para a construção de uma modelagem estatística que permita obter esses dados de uma maneira mais rápida, de forma a prever com precisão os valores de transmitância atmosférica correspondentes ao espectro infravermelho de faixa longa (LWIR do inglês *Long Wave Infrared*). Isso se justifica uma vez que a faixa LWIR foi escolhida por ser a faixa operada pelo IRST do Gripen.

A princípio, a abordagem escolhida para esse desafio foi a regressão linear múltipla. No entanto, essa estratégia não se mostrou viável, conforme revelado pela análise dos resíduos e pela incapacidade do método em limitar os valores previstos no intervalo de 0 a 1, o qual é próprio dos valores de transmitância. Com isso em mente, foi realizada uma tentativa com a regressão beta, com o intuito de manter os valores dentro deste intervalo. Todavia, os valores preditos se mostraram significativamente diferentes do esperado, o que indicou que um ajuste eficaz das variáveis não estava sendo alcançado. Diante dessas dificuldades, a decisão foi recorrer ao aprendizado de máquina, utilizando o método de bagging de Random Forest. Esta abordagem, conforme destacado por [6], já se mostrou eficiente em situações semelhantes, o que a tornou uma alternativa promissora para este trabalho.

II. MATERIAIS E METODOS

A. Software MODTRAN

No presente trabalho, utilizou-se o software MODTRAN para a geração de dados de transmitância atmosférica. O MODTRAN é um sistema de modelagem atmosférica que é amplamente reconhecido por sua precisão e confiabilidade em gerar transmitâncias, reflectâncias e radiações em uma ampla gama de comprimentos de onda e condições atmosféricas [7].

O MODTRAN, desenvolvido e mantido pelo *Spectral Sciences, Inc.* e o *Air Force Research Laboratory*, oferece uma maneira prática e eficiente de modelar as propriedades de absorção e dispersão da

José R. C. Neto, joseroddrigues.c@gmail.com; Mauri A. de Oliveira, Mauri@ita.br; Caio A. de M. Silvestre silvestrecamas@fab.mil.br.

atmosfera, considerando várias variáveis que vão desde o perfil atmosférico específico até a altitude e ângulo solar [8].

Neste estudo, o software foi usado para gerar uma série de dados de transmitância para uma variedade de condições atmosféricas. Esses dados foram então utilizados como a base para uma modelagem estatística que visava predizer a transmitância sob várias condições.

B. Obtenção dos Dados

O processo de obtenção desses dados foi gerado em um computador pessoal com processador Intel Core i7-7700HQ de 2,80 gigahertzs e memória RAM de 16 gigabytes, e através dele foi gerada primeiramente uma tabela com 59.614 linhas de dados com 10 variáveis (colunas), que foram:

I. Espectro eletromagnético;
II. Modelo atmosférico;
III. Modelo de aerossol;
IV. Quantidade de CO_2 (em partes por milhão em volume (ppmv));
V. Temperatura do alvo (em Kelvin);
VI. Visibilidade atmosférica (em quilômetros);
VII. Altitude do alvo (em quilômetros);
VIII. Ângulo zenital do alvo (em graus);
IX. Distância do alvo (em quilômetros);
X. Transmitância efetiva.

Neste estudo, o foco foi na região do espectro eletromagnético LWIR, especificamente na faixa de comprimento de onda de 8,0 a 12,0 μm. Como essa variável é constante, ela não foi considerada na previsão da transmitância atmosférica efetiva.

Foram estabelecidos três diferentes modelos atmosféricos para análise: "*Tropical Model*", "*Midlatitude Summer*" e "*Midlatitude Winter*". No ambiente de programação RStudio, esses modelos foram codificados como variáveis categóricas, atribuindo-se os valores 1, 2 e 3, respectivamente.

Consideraram-se também seis diferentes modelos de aerossol: ausência de aerossol, "rural – vis = 23km", "rural – vis = 5km", "Navy Maritime/Maritime – vis = 23km" e "Urban – vis = 5km". Novamente, no RStudio, esses modelos foram codificados como variáveis categóricas, recebendo as designações de 0,1, 2, 3, 4 e 5, respectivamente.

Para as variáveis IV, V, VI, VII, VIII e IX, foram determinados um mínimo de três valores distintos para cada uma. Assim, utilizou-se um código *Python* para combinar todos esses valores, visando calcular o mais amplo espectro possível de respostas para a transmitância efetiva.

Importante enfatizar que os valores produzidos pelo MODTRAN são resultantes da combinação das variáveis em estudo. Contudo, a geração de uma ampla gama de valores distintos para uma única variável mostra-se impraticável, uma vez que, além de ser um processo notavelmente demorado, algumas das variáveis são contínuas. Desta forma, a geração de todas as possíveis transmitâncias torna-se inviável, pois implicaria em uma quantidade infinita de valores.

Prosseguiu-se com a geração de um conjunto adicional de 15.553 dados, designados aqui como dados de validação. Esses dados foram empregados para avaliar a eficácia do modelo de Random Forest. Mesmo que os dados de validação contivessem uma variedade de variáveis distintas, os três tipos de Modelos Atmosféricos e os cinco modelos de Aerossol foram mantidos consistentes com os dados de treinamento.

Adotando essa metodologia, foi possível avaliar a suposição proposta por [9] de que é fundamental manter as características consistentes ao treinar um modelo de *Random Forest* e ao aplicá-lo a novos dados. Caso contrário, o modelo gerado pode não ser válido.

Dividiu-se o primeiro conjunto de dados em treino e teste, comumente praticado em aprendizado de máquina. Uma proporção de 70% para treino e 30% para teste foi utilizada, de acordo com as recomendações de [10]. Essa divisão é essencial para evitar o overfitting e assegurar que o modelo generalize bem para novos dados. Para garantir a reprodutibilidade, aplicou-se a função set.seed com o valor de 770 no ambiente de programação, outra prática recomendada por [10] para garantir resultados consistentes e replicáveis.

C. Aplicação da Regressão Linear Múltipla

A seleção de variáveis é crucial em modelos de regressão linear múltipla para identificar as variáveis independentes mais relevantes. Utilizamos o Critério de Informação de Akaike (AIC) nesse processo. O AIC, concebido por [11] quantifica a perda de informação de um modelo estatístico, favorecendo modelos com menor AIC. A seleção de variáveis baseada em AIC foi realizada utilizando a função 'step' do pacote 'MASS' em R para a seleção backward, removendo iterativamente as variáveis que resultam no menor aumento do AIC.

Nesta análise, descobriu-se que a remoção da variável CO2 resultou no menor valor de AIC, conforme demonstrado na Fig. 1, sugerindo que este modelo simplificado é preferível considerando-se o equilíbrio entre a precisão do modelo (ou seja, quão bem o modelo se ajusta aos dados) e a complexidade do modelo (ou seja, o número de variáveis no modelo).

Backward Elimination Summary

Variable	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
Full Model	-32879.673	1110.484	1204.076	0.52022	0.52007
CO2	-32881.558	1110.487	1204.073	0.52022	0.52008

Fig. 1. AIC para Modelos de Regressão com e sem a Variável CO2.

Primeiramente, elaborou-se um gráfico para visualizar a transmitância dos dados, como pode ser observado na Fig. 2. Esta etapa é essencial na análise de regressão, pois proporciona um entendimento preliminar sobre as relações potenciais entre as variáveis em estudo. A partir deste gráfico, é possível perceber tendências, relações e possíveis outliers que podem influenciar a análise subsequente.

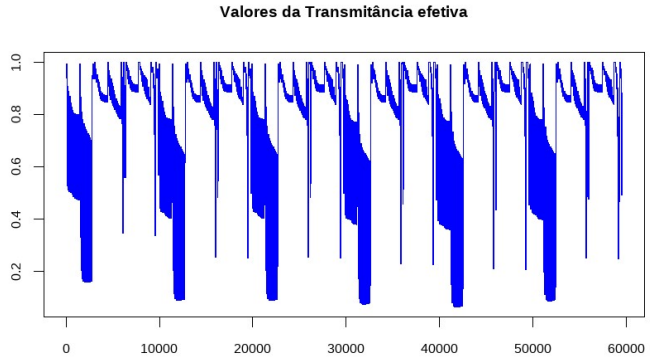

Fig 2 - Gráfico de pontos da Transmitância efetiva. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

Prosseguindo com o trabalho, ocorreu o desenvolvimento de modelos de regressão. Com base em estudos anteriores, já se antecipava que a abordagem de regressão linear múltipla padrão poderia não ser idealmente adequada aos dados em questão. Como resultado, o método AIC foi aplicado, levando à exclusão da variável CO2 na elaboração do modelo_II.

Ao avaliar o desempenho do modelo_II, foi encontrado um somatório dos quadrados dos resíduos de 1110,49, bem como um

coeficiente de determinação (R-quadrado) de 0,52, como ilustrado na Fig. 3.

Model Summary

R	0.721	RMSE	0.163
R-Squared	0.520	Coef. Var	20.995
Adj. R-Squared	0.520	MSE	0.027
Pred R-Squared	0.520	MAE	0.124

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1204.073	12	100.339	3769.573	0.0000
Residual	1110.487	41719	0.027		
Total	2314.560	41731			

Fig. 3. Tabela ANOVA do Modelo II.

Na busca contínua por um modelo mais preciso, foram realizadas várias iterações entre as variáveis após a criação do modelo_II. A adoção dessa abordagem, embora se saiba que ela pode introduzir outros desafios ao modelo, como problemas de normalidade e multicolinearidade, foi motivada pelo desejo de demonstrar que, mesmo com várias alterações, seria difícil criar um modelo que previsse precisamente a transmitância. Além da divergência dos valores, os valores preditos em alguns casos excederam 1, invalidando o modelo, pois os valores de transmitância variam apenas de 0 a 1.

Nesse contexto, foi desenvolvido o modelo_III, que alcançou um coeficiente de determinação (R-quadrado) de 0,63, conforme mostrado na Fig. 16. A formulação de regressão usada para construir o modelo_III foi a seguinte: modelo_III <- lm(Transmitância ~ Modelo + Altitude + Temperatura + Distância + Ângulo + Visibilidade + (Ângulo * Distância) + (Visibilidade * Altitude) + (Modelo * Altitude) + (Altitude * Distância), data = treino)

Model Summary

R	0.796	RMSE	0.143
R-Squared	0.633	Coef. Var	18.361
Adj. R-Squared	0.633	MSE	0.020
Pred R-Squared	0.633	MAE	0.104

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1465.222	12	122.102	5997.577	0.0000
Residual	849.337	41719	0.020		
Total	2314.560	41731			

Fig. 4. Tabela ANOVA do Modelo III.

Depois de uma série de iterações e considerações, foi desenvolvido o modelo_IV <- lm(Transmitancia ~ Modelo + Altitude + Temperatura + Distancia + Angulo + Visibilidade + (Angulo * Distancia) + I(Altitude^2) + I(Distancia^2) + (Visibilidade * Altitude) + (Modelo * Altitude) + (Altitude * Distancia), data = treino)

Apesar deste modelo não cumprir completamente todos os pressupostos da análise de resíduos, conseguiu-se alcançar um coeficiente de determinação (R-quadrado) mais satisfatório de 0,75, como demonstrado na Fig. 17. No entanto, é crucial observar que, mesmo com essa melhora significativa na capacidade de predição, o modelo_IV não pode ser considerado completamente válido para os propósitos estabelecidos.

A formulação do modelo_IV incorporou tanto termos individuais quanto interações entre várias variáveis, tentando capturar efeitos mais complexos presentes no sistema em análise. Contudo, mesmo com todas essas modificações e a incorporação de termos de interação, a previsão precisa da transmitância por meio de um modelo de regressão linear múltipla continuou sendo um desafio.

Model Summary

R	0.867	RMSE	0.117
R-Squared	0.752	Coef. Var	15.104
Adj. R-Squared	0.752	MSE	0.014
Pred R-Squared	0.752	MAE	0.084

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1739.850	14	124.275	9020.873	0.0000
Residual	574.709	41717	0.014		
Total	2314.560	41731			

Fig. 5. Tabela ANOVA do Modelo IV.

D. Aplicação da Regressão Beta

Em face dos desafios encontrados com a regressão linear, principalmente por não conseguir limitar o valor predito da transmitância entre 0 e 1, buscou-se um método alternativo que pudesse satisfazer essa necessidade. Assim, descobriu-se à regressão beta, que possibilita limitar o valor predito entre 0 e 1.

A regressão beta, introduzida por [12], modela respostas seguindo uma distribuição beta, adequada para variáveis contínuas no intervalo $0 < y < 1$. A densidade beta é modelada por (1), com a média e variância dados por (2) e (3) respectivamente.

$$\pi(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \qquad 0 < y < 1, \tag{1}$$

$$E(y) = \frac{p}{(p+q)}, \tag{2}$$

$$var(y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{3}$$

Uma parametrização alternativa introduzida modifica a distribuição para permitir uma estrutura de regressão para a média da variável resposta, resultando em (4) e as novas formas para a média e variância (5) e (6).

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\,\phi)\Gamma\big((1-\mu)\phi\big)} y^{\mu\,\phi-1} \cdot (1 - y)^{(1-\mu)\,\phi-1}, \qquad 0 < y < 1, \tag{4}$$

$$E(y) = \mu, \tag{5}$$

$$var(y) = \frac{V(\mu)}{1+\phi}. \tag{6}$$

O modelo de regressão beta permite estimar os parâmetros desconhecidos através da máxima verossimilhança, usar técnicas de diagnóstico e fazer inferências em grandes amostras [12] também propõem um modelo de regressão para essa distribuição, descrito por (7), (8) e (9).

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = n_{t;} \tag{7}$$

$$l(\beta, \phi) = \sum_{t=1}^{n} l_t(\mu_t, \phi); \tag{8}$$

$$\ell_i(\mu_i, \phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma\big((1-\mu_i)\phi\big) + (\mu_i\phi - 1)\log yi + \{(1-\mu i)\phi - 1\} log(1-yi). \tag{9}$$

Esses parâmetros são obtidos numericamente usando equações derivadas das funções escore, e são realizados testes de hipóteses para avaliar a adequação do modelo

Nesta abordagem, o modelo foi iniciado com todas as variáveis possíveis. Após análise dos resultados, notou-se que a variável CO2 não era estatisticamente significativa para o modelo, levando à decisão de removê-la da análise. Isso resultou na criação do primeiro modelo beta, denominado modelrg_probit.

Vale destacar que a função de ligação probit foi escolhida para esse modelo, pois dentre todas as opções testadas, essa foi a que apresentou o melhor pseudo-R-quadrado.

Para analisar o modelo de regressão beta, o pacote betareg foi utilizado, que oferece várias funções para avaliação e diagnóstico deste tipo de modelo. A função summary() do pacote é particularmente útil, pois fornece uma visão geral abrangente do modelo, embora não produza uma tabela ANOVA completa, como ocorre com os modelos de regressão linear padrão.

Ao executar a função summary() em um objeto betareg, obtém-se uma tabela que apresenta detalhes significativos para cada variável preditora no modelo. Cada linha da tabela corresponde a uma variável diferente e cada coluna apresenta um tipo diferente de informação sobre essa variável.

É importante salientar que a função summary() também fornece o pseudo-R-quadrado, que é uma medida semelhante ao R-quadrado em regressão linear, mas adaptada para modelos de regressão generalizados, como este. Ele dá uma indicação de quão bem o modelo está ajustando os dados, com valores mais próximos de 1 indicando um melhor ajuste.

Ao aplicar a função summary no modelo modelrg_probit, foram obtidos os resultados apresentados na Fig. 5, destacando-se o pseudo-R-Quadrado de 0,5435, um indicador da qualidade do ajuste do modelo beta.

```
Call:
betareg(formula = Transmitancia ~ Modelo + Aerossol + Altitude + Distancia + Angulo +
    Temperatura + Visibilidade, data = treino, link = "probit")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.3191 -0.5356 -0.0092  0.5220  2.9982

Coefficients (mean model with probit link):
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)   6.831e-01  1.145e-02   59.682  < 2e-16 ***
Modelo2       7.533e-02  5.801e-03   12.985  < 2e-16 ***
Modelo3       3.045e-01  5.793e-03   52.566  < 2e-16 ***
Aerossol1    -4.906e-02  8.134e-03   -6.031 1.63e-09 ***
Aerossol2    -5.125e-02  8.144e-03   -6.294 3.10e-10 ***
Aerossol3    -7.222e-02  8.149e-03   -8.862  < 2e-16 ***
Aerossol4    -7.598e-02  8.141e-03   -9.333  < 2e-16 ***
Aerossol5    -5.882e-02  8.150e-03   -7.217 5.31e-13 ***
Altitude      1.258e-01  7.527e-04  167.094  < 2e-16 ***
Distancia    -1.952e-02  1.675e-04 -116.540  < 2e-16 ***
Angulo       -1.733e-03  6.619e-05  -26.178  < 2e-16 ***
Temperatura  -3.130e-05  5.274e-06   -5.935 2.94e-09 ***
Visibilidade  4.128e-03  2.883e-04   14.317  < 2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  5.87051    0.04127   142.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 3.697e+04 on 14 Df
Pseudo R-squared: 0.5435
Number of iterations: 22 (BFGS) + 1 (Fisher scoring)
```

Fig. 6. Summary do Modelrg_probit.

Embora a regressão beta tenha resolvido um dos principais problemas encontrados com a Regressão Linear Múltipla (RLM), isto é, a limitação dos valores preditos entre 0 e 1, notou-se que os valores preditos ainda se distanciavam significativamente do esperado. Assim, decidiu-se explorar novas transformações nas variáveis com o intuito de melhorar o pseudo-R-Quadrado, mesmo sabendo que essa estratégia poderia levar à violação de alguns pressupostos da análise de resíduos.

A partir dessa análise, chegou-se a um modelo que se assemelhava ao melhor modelo predito pela RLM. De maneira semelhante, o melhor link escolhido para este novo modelo foi o probit.

Este modelo resultou em um Pseudo-R-Quadrado de 0,7641, representando uma melhoria significativa na qualidade de ajuste em comparação com o modelo anterior, como pode ser observado na Fig. 6.

```
Call:
betareg(formula = Transmitancia ~ Modelo + Altitude + Temperatura + Distancia +
    Angulo + Visibilidade + (Angulo * Distancia) + I(Altitude^2) + I(Distancia^2) +
    (Visibilidade * Altitude) + (Modelo * Altitude) + (Altitude * Distancia), data = treino,
    link = "probit")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.9169 -0.4509 -0.0046  0.5777  3.3614

Coefficients (mean model with probit link):
                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)           5.489e-01  1.117e-02   49.151   <2e-16 ***
Modelo2               2.635e-01  7.035e-03   37.450   <2e-16 ***
Modelo3               8.185e-01  7.158e-03  114.345   <2e-16 ***
Altitude              3.973e-01  2.574e-03  154.369   <2e-16 ***
Temperatura          -3.435e-05  4.094e-06   -8.392   <2e-16 ***
Distancia            -6.345e-02  5.966e-04 -106.355   <2e-16 ***
Angulo               -1.192e-03  7.281e-05  -16.376   <2e-16 ***
Visibilidade          1.106e-02  3.549e-04   31.168   <2e-16 ***
I(Altitude^2)        -2.464e-02  2.468e-04  -99.853   <2e-16 ***
I(Distancia^2)        1.162e-03  9.771e-06  118.913   <2e-16 ***
Distancia:Angulo     -2.575e-04  3.939e-06  -65.373   <2e-16 ***
Altitude:Visibilidade -1.502e-03  6.823e-05  -22.010   <2e-16 ***
Modelo2:Altitude     -4.559e-02  1.370e-03  -33.289   <2e-16 ***
Modelo3:Altitude     -1.171e-01  1.380e-03  -84.838   <2e-16 ***
Altitude:Distancia    1.141e-03  3.732e-05   30.581   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi) 12.79584    0.09084   140.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 5.192e+04 on 16 Df
Pseudo R-squared: 0.7643
Number of iterations: 26 (BFGS) + 2 (Fisher scoring)
```

Fig. 7. Summary do Modelrg2_probit.

E. Aplicação do Random Forest

Face às dificuldades encontradas na modelagem através da Regressão Linear Múltipla e da Regressão Beta, a abordagem do Random Forest foi explorada para a predição da transmitância atmosférica. O processo iniciou com um estudo para determinar o número ideal de árvores para o modelo de *Random Forest*. Utilizou-se um loop for para criar modelos com diferentes quantidades de árvores, especificamente 50, 100, 150, 200 e 400.

Simultaneamente, a otimização do número de variáveis no modelo foi tentada, variando entre 1 e o número máximo de variáveis disponíveis nos dados. Desta forma, buscou-se a quantidade de variáveis que proporcionaria o melhor desempenho para o modelo. Para avaliar o desempenho de cada configuração o *Out-of-Bag* (OOB) foi utilizado como métrica de referência.

Notou-se que a partir de 150 árvores, o valor de OOB se torna praticamente constante, indicando essa como a quantidade adequada de árvores para o modelo de *Random Forest*.

Após estabelecer o número apropriado de árvores, a atenção se voltou para a otimização do número de variáveis. Através da análise do coeficiente de determinação (R-quadrado) para cada configuração de variáveis, determinou-se que a inclusão de 7 variáveis proporcionava o melhor desempenho para o modelo. Isso levou à definição da estrutura do modelo de *Random Forest*, baseado em 150 árvores e 7 variáveis.

Finalmente, para a geração do modelo de *Random Forest*, a função 'train' do pacote 'caret' foi utilizada. Essa função é altamente versátil e permite a implementação de uma variedade de modelos preditivos. Especificamente para essa análise, o método 'rf', que se refere ao Random Forest, foi selecionado.

Optou-se por uma validação cruzada de 10-fold com a construção de 150 árvores e a inclusão das 7 variáveis, conforme definido anteriormente.

III. RESULTADOS E DISCUÇÕES

A. Resultados da Regressão Linear Múltipla

Conforme discutido, foram enfrentados desafios significativos ao tentar aplicar a regressão linear múltipla para prever a transmitância atmosférica, principalmente devido ao seu caráter limitado ao intervalo de 0 a 1. Essa característica impôs restrições notáveis ao uso da regressão linear múltipla, uma vez que este método estatístico não possui mecanismos para impor limites superiores ou inferiores aos valores preditos. Foram apresentados diversos modelos e transformações na tentativa de maximizar o coeficiente de determinação R-quadrado, mesmo cientes de que tais modificações não seriam suficientes para evitar a ultrapassagem do limite de 1 nas previsões.

Ao aplicar o modelo_IV na base de teste, o objetivo primordial foi criar um gráfico comparativo dos valores preditos versus os valores reais da transmitância atmosférica. Essa iniciativa visava destacar o problema de ultrapassagem do valor de 1 nas previsões, uma dificuldade inerente à utilização da regressão linear múltipla neste contexto. A Fig. 7 demonstra o resultado dessa comparação.

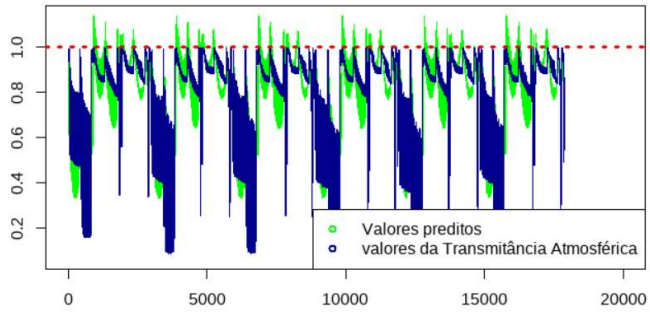

Fig. 8. Comparação entre os valores preditos pela RLM e a Transmitância. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

B. Resultados da Regressão Beta

Em resposta ao desafio de limitar os valores preditos entre 0 e 1, a escolha foi pela regressão beta. Como detalhado na Seção 3.6.2, foi selecionado um modelo que exibia o melhor Pseudo-R-Quadrado. No entanto, mesmo com a implementação da regressão beta, que realmente limitou os valores preditos entre 0 e 1, ainda se enfrentou um erro considerável na predição dos valores. Essa discrepância é ilustrada na Fig. 8.

É importante destacar que o gráfico foi gerado utilizando a base de testes para examinar os valores preditos pelo modelo em questão. Mesmo com a melhoria proporcionada pela regressão beta, ainda existem desafios significativos a serem superados para aprimorar a precisão do modelo de previsão.

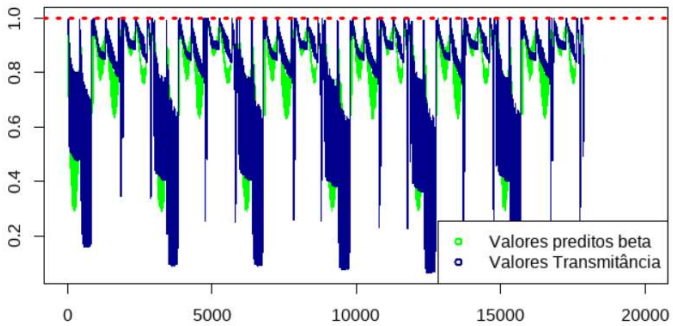

Fig. 9. Comparação entre os valores preditos pela Regressão Beta e a Transmitância. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

C. Resultados do Random Forest

Finalmente, após diversas iterações e testes, chegou-se ao modelo de *Random Forest*. O modelo final escolhido baseou-se em 150 árvores e 7 variáveis. Notavelmente, este modelo produziu um coeficiente de determinação R-quadrado impressionante de 99,99% quando aplicado à base de dados de teste.

A precisão deste modelo é claramente demonstrada na Fig. 9. A maioria dos valores preditos coincidem com os valores reais da transmitância, deixando pouco ou nenhum espaço visível na cor vermelha. Isso sugere que quase todas as previsões foram precisamente calculadas, indicando um ajuste extremamente bom do modelo aos dados. Em outras palavras, o modelo de Random Forest demonstrou uma habilidade notável para prever a transmitância atmosférica com alta precisão.

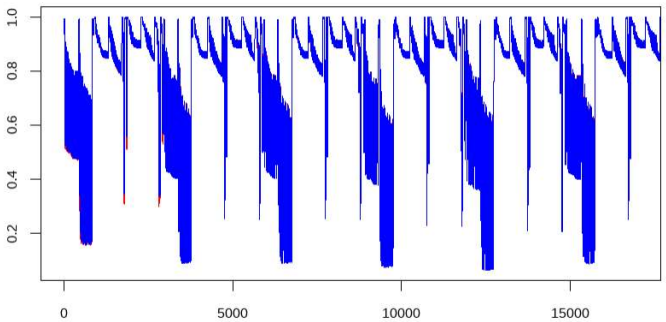

Fig. 10. Comparação entre os valores preditos pelo *Random Forest* e a Transmitância. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

Para fornecer uma análise ainda mais detalhada da precisão do modelo, decidiu-se ampliar os resultados, focando em uma seleção menor de dados. Na Fig. 10, é visualizada uma segmentação de 300 valores preditos. Isso permitiu examinar mais de perto a correspondência entre os valores previstos pelo modelo de Random Forest e os valores reais de transmitância atmosférica. Essa perspectiva ampliada oferece uma evidência mais granular do alto grau de precisão do modelo.

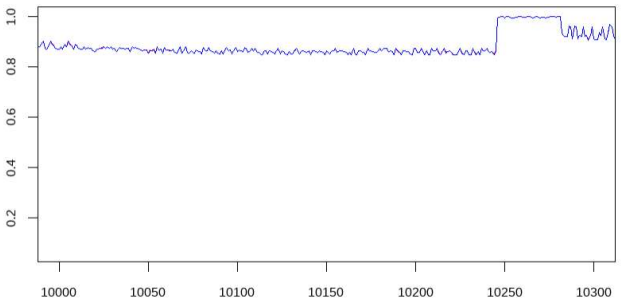

Fig. 11. Comparação entre apenas alguns valores preditos no algoritmo e os observados. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

Embora o modelo de *Random Forest* tenha demonstrado um desempenho impressionante com a base de testes, é importante salientar que esta base foi extraída do mesmo conjunto de dados original utilizado para treinamento. Conforme discutido, o MODTRAN gera os valores da transmitância por meio de combinações de variáveis, resultando em uma ampla gama de resultados. No entanto, devido ao volume massivo dessas combinações, o conjunto de dados não apresenta uma diversidade abrangente em todas as variáveis. Por exemplo, a temperatura, apesar

de ser uma variável contínua, só possui quatro valores distintos no conjunto de dados.

O *Random Forest* é reconhecido por sua eficácia na previsão de resultados quando os valores utilizados para o treinamento estão presentes no conjunto de dados. Contudo, ao aplicar o modelo treinado em um banco de dados distinto, com variações ligeiramente diferentes nos valores das variáveis, a precisão na previsão da transmitância revelou-se insatisfatória, apresentando um R-quadrado de apenas 50,90%. Esta situação pode ser claramente visualizada no gráfico apresentado na Fig. 11.

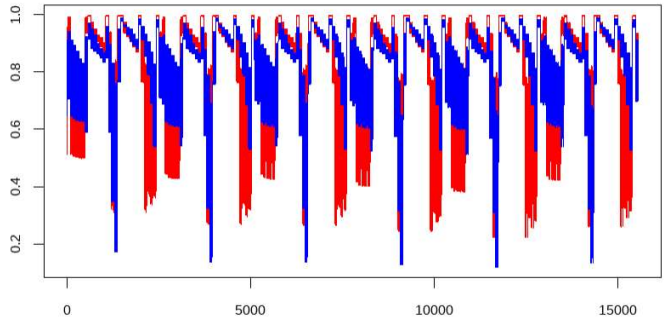

Fig. 12. Comparação entre os valores preditos pelo *Random Forest* usando uma base de dados de validação e a Transmitância. Eixo x denota o número sequencial de medições, e o eixo y mostra a transmitância efetiva para cada medição.

IV. CONCLUSÕES

No contexto militar contemporâneo, a análise da transmitância atmosférica tem se tornado cada vez mais essencial, especialmente devido ao aprimoramento contínuo dos armamentos. Este fenômeno é particularmente notável no Brasil, marcado pela implementação do Gripen E/F e da tecnologia IRST. Como exemplo, existem modelos que incorporam a transmitância em um algoritmo voltado para o planejamento de missões aéreas que enfrentam ameaças no espectro infravermelho. Esta aplicação realça a importância deste campo de estudo na esfera militar atual.

No entanto, a utilização do MODTRAN, software pago para o cálculo da transmitância, representa um obstáculo para o estudo mais amplo. Por este motivo, a alternativa explorada foi a aplicação de modelos estatísticos e de aprendizado de máquina para prever a transmitância atmosférica de maneira mais eficiente.

Ainda assim, ao longo deste estudo, a aplicação da regressão linear como ferramenta de previsão mostrou-se inviável, principalmente por não conseguir limitar os valores previstos entre 0 e 1, além de apresentar diferenças significativas entre os valores preditos e os observados. Nesse contexto, a regressão beta emergiu como uma alternativa, sendo um modelo de regressão desenvolvido por pesquisadores brasileiros, capaz de limitar o valor previsto entre 0 e 1, sendo amplamente empregado na previsão de porcentagens.

Mesmo assim, apesar da regressão beta resolver a limitação da regressão linear, ainda se constataram diferenças consideráveis entre os valores previstos e os observados da transmitância. Embora o método tenha aprimorado a extensão dos resultados, o desafio de alinhar os valores preditos aos observados persistiu.

Finalmente, recorreu-se ao uso do aprendizado de máquina, especificamente o método *Random Forest*, que exibiu excelentes resultados ao prever um R-quadrado de 99,99%. No entanto, como é sabido, o *Random Forest* requer que os valores das variáveis que foram treinados estejam presentes no novo conjunto de dados a ser previsto. Ao aplicar o modelo a um novo conjunto de dados com variáveis ligeiramente diferentes das usadas para treiná-lo, verificou-se que a qualidade da previsão diminuiu significativamente, apresentando um R-quadrado de 50,9%.

Isto permite concluir que o uso do *Random Forest* é muito eficaz para a predição dos valores de transmitância, porém, sua implementação em um software exigiria um treinamento com uma variedade muito maior de dados. Este é um desafio de grande magnitude, pois gerar todos esses dados com os recursos atualmente disponíveis seria impraticável. Além disso, seria necessária uma infraestrutura computacional robusta para processar e gerar tal modelo, pois estaria lidando com uma quantidade massiva de dados, na ordem de milhões.

REFERÊNCIAS

[1] GAITANAKIS, G. K. *et al*. Infrared search & track systems as anti-stealth approach**.** Journal of Computations Modelling, v. 9, n. 1, p. 33-53, 2019.

[2] SILVA, Mauricio de Almeida Vellasquez. Modelo de predição de alcance de sistemas Infrared Search and Track. 2021. 61f. Trabalho de Conclusão de Curso em Guerra Eletrônica (Especialização em Análise de Ambiente Eletromagnético) – Instituto Tecnológico de Aeronáutica, São José dos Campos, 2021.

[3] ARAUJO DE CARVALHO, Diego; Análise da influência da radiação refletida do sol no cálculo da assinatura infravermelha de aeronave nas bandas espectrais do infravermelho. 2019. 55f. Trabalho de Conclusão de Curso. (Lato Sensu) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

[4] DANTAS, Joao P. A.; COSTA, Andre N.; GERALDO, Diego; MAXIMO, Marcos R. O. A.; YONEYAMA, Takashi. Engagement Decision Support for Beyond Visual Range Air Combat. In: LATIN AMERICAN ROBOTICS SYMPOSIUM (LARS), 21.; BRAZILIAN SYMPOSIUM ON ROBOTICS (SBR), 2021; WORKSHOP ON ROBOTICS IN EDUCATION (WRE), 2021. Anais [...]. [S.l.]: IEEE, 2021. DOI: 10.1109/LARS/SBR/WRE54079.2021.9605380.

[5] LIMA FILHO, Geraldo Mulato de; KUROSWISKI, André Rossi; MEDEIROS, Felipe Leonardo Lôbo; VOSKUIJL, Mark; MONSUUR, Herman; PASSARO, Angelo. Optimization of Unmanned Air Vehicle Tactical Formation in War Games. **IEEE Access**, v. 10, n. 1, p. 1-20, 18 fev. 2022. DOI: 10.1109/ACCESS.2022.3152768.

[6] PEIXOTO JÚNIOR, Paulo José Rovéglia. Modelagem estatística da transmitância atmosférica através de regressão linear e aprendizado de máquina. 2022. 89f. Trabalho de Conclusão de Curso. (Lato Sensu) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

[7] BERK, A. et al. MODTRAN 4 version 2 user's manual. Hanscom: Air Force Research Laboratory Space Vehicles Directorate, 1999.

[8] BERK, A.; BERNSTEIN, L.S.; ROBERTSON, D.C. MODTRAN: A moderate resolution model for LOWTRAN 7. In: GL-TM-89-0122, 2006. Anais... [S.l.: s.n.], 2006.

[9] BREIMAN, L. Random forests. Machine Learning, v.45, n. 1, p. 5–32. 2001.

[10] HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2009.

[11] AKAIKE, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, v. 19, n. 6, p. 716-723, 1974.

[12] FERRARI, S. L. P.; CRIBARI-NETO, F. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, v. 31, n. 7, p. 799-815, 2004.