

Bolsista: Enzo Laragnoit Fernandes

Previsão de Irradiação Solar para Sistemas Fotovoltaicos utilizando *Machine Learning*

**Uma abordagem de Aprendizagem Supervisionada para Previsão de
Irradiação Solar no Estado de São Paulo**

FAPESP

Fundação de Amparo à Pesquisa do Estado de São Paulo

Processo: 2020/09607-9

Vigência: 01/09/2020 a 31/08/2021

Relatório Científico Final

Período: 11/02/2021 a 31/08/2021

Setembro de 2021

1 Resumo do Projeto Proposto

Um dos principais desafios para o século XXI é conciliar a busca por soluções energéticas com o desenvolvimento sustentável sem causar maiores prejuízos ao já afetado meio ambiente. Com a crescente demanda por energia elétrica (EPE, 2017), movida pela progressiva informatização da sociedade moderna, surgem diferentes formas renováveis de geração de energia elétrica em alternativa aos finitos recursos não renováveis como carvão mineral e petróleo.

Uma das alternativas promissoras é a conversão de energia solar em energia elétrica por meio de painéis solares fotovoltaicos que transformam a irradiação solar em energia elétrica que pode ser prontamente utilizada ou armazenada em baterias sob a forma de energia química. Presente em grande abundância e em praticamente todos os lugares, a quantidade de energia incidente na Terra proveniente do Sol é aproximadamente 10 000 vezes maior que a quantidade de energia consumida pela humanidade (SMETS et al., 2015). A energia solar apresenta vantagens quando comparada a outras formas de energia renovável, como a eólica e a hidrelétrica, pela facilidade de instalação, considerando aspectos de construção, danos ambientais, custos operacionais e manutenção. Além disso, apresenta a possibilidade de ser utilizada sem grandes dificuldades nos ambientes urbanos e também oferece como alternativa um modelo de geração de energia descentralizada, no qual consumidores podem gerar sua própria energia. O aumento da eficiência energética e a constante redução dos custos dos sistemas fotovoltaicos justificam sua tendência de crescimento e inserção nas matrizes energéticas (ACHILLES, 2013).

Outra área de grandes avanços e crescentes expectativas é a de Inteligência Artificial, especialmente a subárea de Aprendizado de Máquina (PERRAULT et al., 2019). Possibilitada pelo aprimoramento das técnicas e pelo aumento do poder computacional dos sistemas modernos, sua presença em aplicações utilizadas no cotidiano como reconhecimento de imagens, processamento de linguagem natural, *chatbots* e sistemas de recomendação tornou-se praticamente universal em softwares nas mais diversas plataformas. Grande esperança é depositada nos *insights* e resultados obtidos da aplicação dessas técnicas para trabalhar com questões na fronteira do conhecimento de tópicos fundamentais como energia, saúde e economia em consonância com a computação científica.

Dentre as aplicações possibilitadas pelo uso das técnicas de Aprendizado de Máquina destaca-se, no contexto deste projeto, o uso de Aprendizado de Máquina Supervisionado (*Supervised Machine Learning*) na competência de aferir previsões a partir de um conjunto de dados já conhecidos *a priori*. Essa tarefa é realizada utilizando diferentes algoritmos como em (HAYKIN, 1994; SUYKENS; VANDEWALLE, 1999), cujo propósito geral é obter uma função $f : X \rightarrow Y$ capaz de realizar o mapeamento entre dados de entrada, chamados de atributos (*features*), e as variáveis alvo (*targets*) cujos valores se objetiva determinar. Para este projeto, as informações relacionadas aos dados meteorológicos serão utilizadas como os dados de entrada X da função f e a radiação solar horizontal global será a variável alvo (*target variable*) a ser determinada.

O projeto foi baseado em aplicações similares dessas técnicas em diferentes contextos e localidades como (ALZAHRANI et al., 2017; LI et al., 2016; ASSOULINE; MOHAJERI; SCARTEZZINI, 2017; LOU et al., 2016; ZENG; QIAO, 2013), na intersecção entre as áreas de Inteligência Artificial e Energia Solar.

2 Objetivos e Cronograma

O objetivo proposto nesse projeto foi o de utilizar técnicas de Aprendizado de Máquina a fim de implementar modelos capazes de prever a irradiação solar, medida em kJ/m^2 , em intervalos de antecedência de 60 minutos em pontos específicos no estado de São Paulo. Para tanto, foram estipuladas as metas abaixo e cujo cronograma é apresentado na Tabela 1.

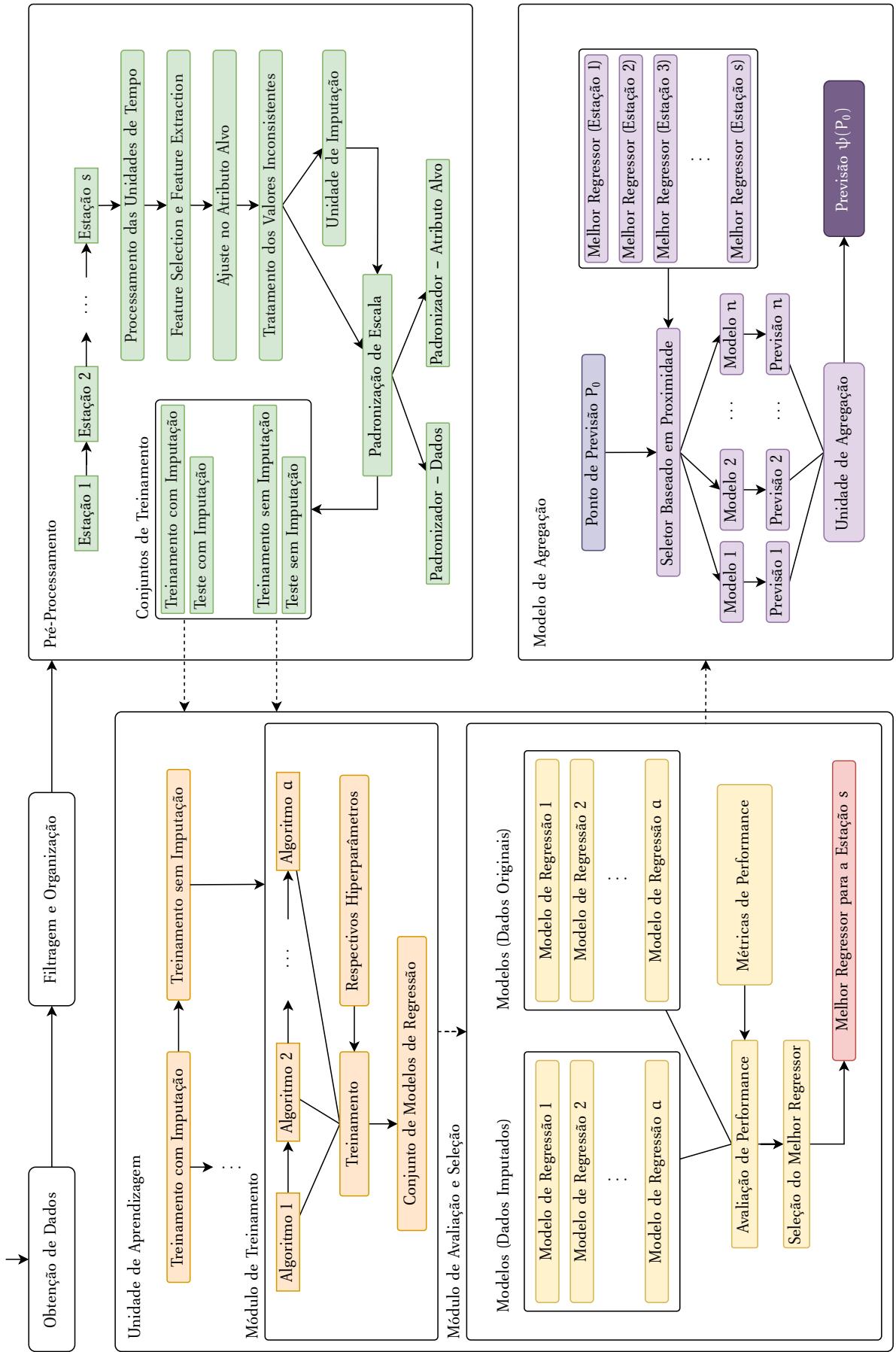
1. Obtenção do conjunto de dados;
2. Geração do conjunto de dados preprocessados;
3. Seleção dos algoritmos para a implementação dos modelos;
4. Implementação dos modelos de previsão a partir de diferentes algoritmos de aprendizagem supervisionada selecionados;
5. Teste e avaliação do desempenho dos modelos em relação ao conjunto de dados obtidos *a priori*;
6. Ajuste dos modelos com base nas informações do item 5;
7. Teste e avaliação do desempenho dos modelos na previsão de irradiação nos locais específicos durante 1 mês;
8. Elaboração dos Relatórios;

Tabela 1 – Cronograma de execução do projeto. As metas em marcadas em verde foram concluídas e compõem este Relatório Científico Final, enquanto as em azul correspondem às metas desenvolvidas durante o período referente ao Relatório Científico anterior.

Metas	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
1												
2												
3												
4												
5												
6												
7												
8												

Fonte: Elaborada pelo autor.

Figura 1 – Esquema de transformações realizadas ao longo da execução do Projeto.



3 Ajustes no Pré Processamento dos Dados

Antes de prosseguir com o treinamento e ajuste dos hiperparâmetros dos modelos, foram realizadas etapas adicionais de pré-processamento do conjunto de dados, no intuito de corrigir inconsistências que não foram observadas nos passos anteriores.

3.1 Conversões de Formato para Unidades de Tempo

Ao longo dos anos foram introduzidas mudanças no formato de registro das variáveis DATA e HORA em cada estação. O método automatizado de leitura dos arquivos em disco considerava os valores nos formatos diferentes com diferentes precisões, resultando no aumento de processamento necessário para inferir os formatos dos atributos e no aumento do espaço necessário para o armazenamento das informações em memória. Para contornar este problema os atributos de tempo foram padronizados, melhorando consideravelmente o tempo de execução das rotinas relacionadas ao processamento destes atributos.

3.2 Duplicidade em Atributos de Localização Geográfica

Na maioria dos pontos de coleta de dados houve dois ou mais valores relativos à localização geográfica — latitude, longitude e altura — da estação. Acredita-se que estas inconsistências ocorreram ou pelo deslocamento físico dos pontos de coleta de dados ao longo dos anos de observação, ou pelo reajuste da precisão destas coordenadas nos *softwares* utilizados para seu registro. Para eliminar as duplicidades nos atributos LATITUDE, LONGITUDE e ALTURA foram escolhidos, para cada estação, a combinação de latitude, longitude e altura mais frequente, ou seja, a combinação que fora registrada durante mais tempo, sendo realizada apenas a substituição dos valores sem que fosse necessário descartar amostras.

3.3 Imputação nos Atributos de Interesse

Em virtude de problemas enfrentados na coleta de dados meteorológicos — *e.g.* exposição ininterrupta às condições climáticas, problemas relacionados a *hardware*, versionamento dos *softwares* nas pontos de coleta de dados, *etc.* — houve perda sensível de informações nos pontos de coleta de dados. Como esta perda não se dava de maneira uniforme entre as estações selecionadas, a estratégia adotada na tentativa de preencher os valores faltantes foi utilizar a técnica de interpolação espacial denominada *Inverse Distance Weighting* de maneira análoga a (KESKIN et al., 2015; OZTURK; KILIC, 2016), de modo que os valores faltantes pudessem ser calculados, quando possível, por meio dos valores dos seus pontos de referência próximos nos mesmos instantes de tempo. O Algoritmo 1 ilustra o procedimento de imputação para um atributo feature em uma

estaçao s_0 em relação a um conjunto de estações selecionadas S — neste contexto, todas as estações utilizadas no projeto.

Algoritmo 1 IDWInputer($s_0, S, \text{feature}, \text{dist_thresh}, \text{min_stations}$)

```

1: neaby_stations  $\leftarrow \emptyset$ 
2: for  $s \in S$  do
3:   if  $\text{distance}(s, s_0) \leq \text{dist\_thresh}$  then
4:      $\text{nearby\_station} \cup \{s\}$ 
5:   end if
6: end for
7: if  $|\text{nearby\_stations}| \geq \text{min\_stations}$  then
8:   for  $t \in \{\min(s_0.\text{timestamp}), \dots, \max(s_0.\text{timestamp})\}$  do
9:     if  $s_0.\text{feature}$  at  $t$  is NA then
10:       $s_0.\text{feature}$  at  $t \leftarrow \text{interpolate}(s_0, \text{nearby\_stations})$ 
11:    end if
12:   end for
13: end if
```

Para o limiar de distância dist_thresh foi adotado o valor de 150 km, enquanto a quantidade mínima de pontos de referência min_estacoes foi de 3 estações, ou seja, ao menos 3 pontos de referência para o cálculo do valor faltante. A aplicação desta técnica resultou em um aumento considerável no número de amostras à disposição para treinamento. A Tabela 4 no Apêndice A apresenta detalhes sobre o resultado da imputação utilizando a técnica de interpolação espacial descrita.

3.4 Tratamento de Valores Faltantes e Inconsistentes

Para os valores faltantes e inconsistentes a estratégia adotada foi a de tirar proveito da implementação da imputação de modo que todos os valores faltantes ou inconsistentes no conjunto de dados fossem substituídos por valores NA, a serem imputados conforme a Seção 3.3. A estratégia adotada anteriormente consistiu em remover todas as amostras do conjunto de dados que contivessem ao menos um valor NA, resultando em grande perda de informações; entretanto parte destas informações pode ser recuperada.

4 Ajuste dos Modelos de Predição

Retomando o contexto da seleção dos algoritmos de aprendizagem supervisionada conforme iniciado na Seção 3.3 do Relatório Científico de Progresso, foram escolhidos os algoritmos *Support Vector Machine*, Redes Neurais Profundas, *Random Forests*, *Extra Trees* e *Extreme Gradient Boosting*.

Para realizar os ajustes dos hiperparâmetros dos modelos foram utilizadas implementações de técnicas como *Grid Search* e *Randomized Search* (BERGSTRA; BENGIO, 2012), disponíveis na

biblioteca Scikit-Learn, para realizar a busca (PEDREGOSA et al., 2011). Como boa prática foi utilizada a validação cruzada com 5 dobras (REFAEILZADEH; TANG; LIU, 2009) apenas no conjunto de dados de treinamento, preservando o conjunto de testes para a avaliação.

A Tabela 2 apresenta os valores dos hiperparâmetros testados durante as rotinas de *Grid Search* para cada um dos algoritmos. Estes valores foram testados em cada um dos modelos e para cada uma das localidades utilizadas no treinamento; uma consequência disso foi que nem todas as estações treinadas com o mesmo algoritmo apresentaram o mesmo conjunto de hiperparâmetros ótimos. Estas diferenças devem-se às características intrínsecas de cada localidade, refletidas em padrões contidos nas observações dos conjuntos de treinamento.

Tabela 2 – Hiperparâmetros testados em cada algoritmo de treinamento.

Algoritmo	Hiperparâmetro	Valores Testados
Support Vector Machine ¹	<code>epsilon</code>	0.1, 0.15, 0.2, 0.4
	<code>gamma</code>	<code>scale, auto</code>
	<code>C</code>	1.0, 2.0, 5.0
Random Forests ²	<code>min_samples_split</code>	2, 20, 100, 250, 500
	<code>min_samples_leaf</code>	1, 10, 50, 150, 500
	<code>max_features</code>	<code>auto, sqrt</code>
	<code>n_estimators</code>	100, 150, 200, 400
Extra Trees ³	<code>min_samples_split</code>	2, 20, 100, 250, 500
	<code>min_samples_leaf</code>	1, 10, 50, 150, 500
	<code>max_features</code>	<code>auto, sqrt</code>
	<code>n_estimators</code>	100, 150, 200, 400
Extreme Gradient Boosting ⁴	<code>eta</code>	0.3, 0.1, 0.5
	<code>gamma</code>	0, 2, 5, 10
	<code>max_depth</code>	6, 8, 10
	<code>sampling_method</code>	<code>uniform, gradient_based</code>
Redes Neurais ⁵	<code>solver</code>	<code>adam, lbfgs, sgd</code>
	<code>activation</code>	<code>relu, logistic</code>
	<code>learning_rate</code>	<code>constant, adaptative</code>

Para as Redes Neurais, além dos hiperparâmetros mostrados na Tabela 2, também foram configurados valores tais como `n_iter_no_change` (número máximo de épocas sem que haja uma alteração no desempenho), `tol` maior que 0,001 e número máximo de iterações (épocas) sobre o conjunto de dados `max_iter = 2000`. Para esta classe de algoritmos em específico

¹ <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

² <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

³ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html>

⁴ <https://xgboost.readthedocs.io/en/latest/parameter.html>

⁵ https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

também foram utilizadas as seguintes arquiteturas: (100,), (200,), (100,100), (150,150), (50,50,20) e (30,30,15); cada arquitetura é representada por uma n -tupla onde n representa o número de camadas ocultas e cada elemento representa a quantidade de neurônios de cada camada oculta.

4.1 Descarte do Uso de Redes Neurais Recorrentes

Apesar de as Redes Neurais Recorrentes possuírem excelentes características para processamento de séries temporais, os modelos obtidos resultantes do treinamento não apresentaram uma capacidade de generalização adequada aos dados e, portanto, foram descartados.

4.2 Modelo Baseado em Ensemble

Para cada localidade de treinamento s foi utilizado um modelo baseado em composição (*ensemble model*) denominado *Stacking Regressor* (WOLPERT, 1992). Este modelo utiliza os demais modelos treinados como blocos de construção para um único estimador de modo que as previsões realizadas por cada um dos estimadores que o compõem são avaliadas por um modelo denominado *meta learner* que determina a melhor maneira de agregar os valores propostos a fim de gerar as previsões finais da composição. Como *meta-learner* foi utilizada uma rede neural com duas camadas ocultas, cada qual com 30 neurônios; neste modelo em particular também foi configurado o hiperparâmetro `max_iter = 1000`.

4.3 Esquema de Execução das Rotinas de Treinamento

A sequência de execução dos algoritmos de treinamento após a definição os hiperparâmetros de busca e da concepção do modelo de agregação intra-local é mostrada seguir:

1. Treinamento dos modelos sem busca de hiperparâmetros;
2. Treinamento do modelo *Stacking Regressor* utilizando modelos sem busca de hiperparâmetros;
3. Treinamento dos modelos com busca de hiperparâmetros;
4. Treinamento do modelo *Stacking Regressor* utilizando modelos com os hiperparâmetros resultados da busca;

Após a realização destes procedimentos foram obtidos 12 modelos para cada local de treinamento s , 5 deles obtidos sem busca de hiperparâmetros (item 1) com adição de 1 modelo *Stacking Regressor* (item 2), e 5 obtidos com a busca de hiperâmetros (item 3) acrescidos de 1 modelo *Stacking Regressor* agregando os modelos após a busca de hiperparâmetros (item 4). Em seguida foi escolhido, dentre os 12 estimadores, o modelo com melhor desempenho.

4.4 Seleção dos Melhores Modelos

Para a seleção do melhor modelo do local s , todos os 12 estimadores foram ordenados em ordem crescente pelos valores das métricas RMSE, MAE, MBE e R^2 , sendo selecionado o primeiro modelo da ordenação resultante. Considerando o contexto do pré-processamento dos dados utilizando a Interpolação Espacial, os passos descritos na Seção 4.3 foram utilizados simultaneamente em dois conjuntos de dados diferentes sendo eles: 1 – conjunto de dados original; e 2 – conjunto de dados com imputação por meio da Interpolação Espacial. Desta forma os modelos treinados utilizando os conjuntos de dados com imputação foram utilizados apenas quando o resultado da aplicação da interpolação espacial resultou em melhor capacidade de generalização.

5 Resultados

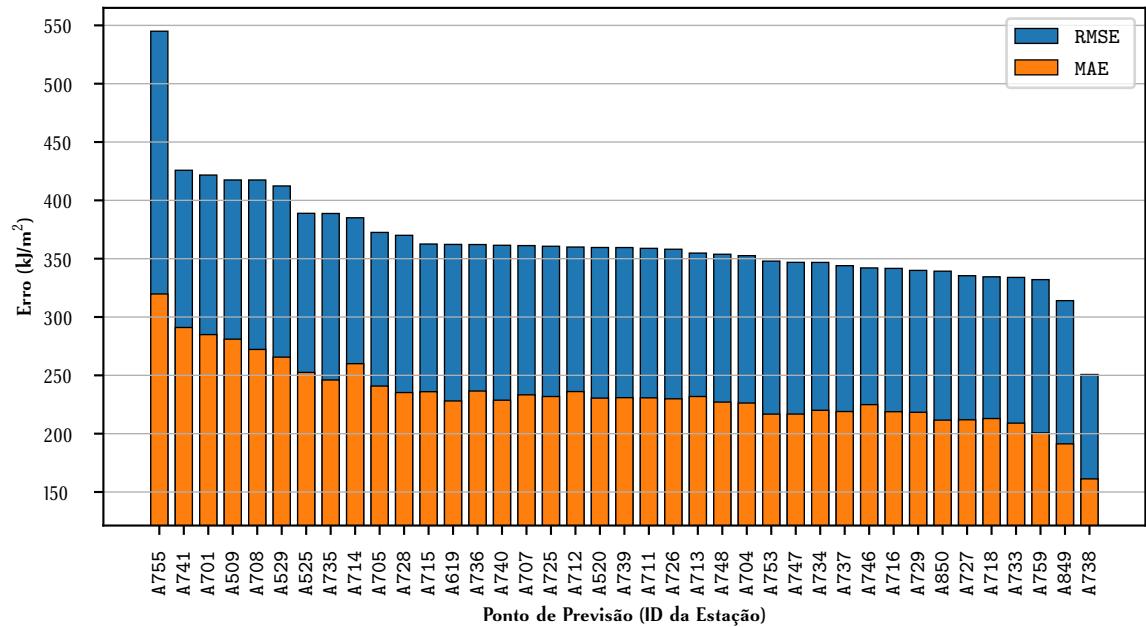
A seguir são apresentados os resultados da implementação dos modelos de previsão. Na Seção 5.1 são apresentados os resultados e a avaliação de desempenho individual dos modelos para cada local de treinamento dos modelos individuais; já a Seção 5.2 apresenta os resultados e a avaliação do desempenho do modelo de agregação proposto, bem como de um modelo de agregação alternativo proposto no período referente a este Relatório Científico.

5.1 Avaliação do Desempenho dos Modelos Individuais

Para realizar a avaliação dos modelos foram utilizados dados com observações contínuas entre os anos de 2018 à 2020 (que fora excluído do conjunto de treinamento) de cada um dos locais selecionados para treinamento.

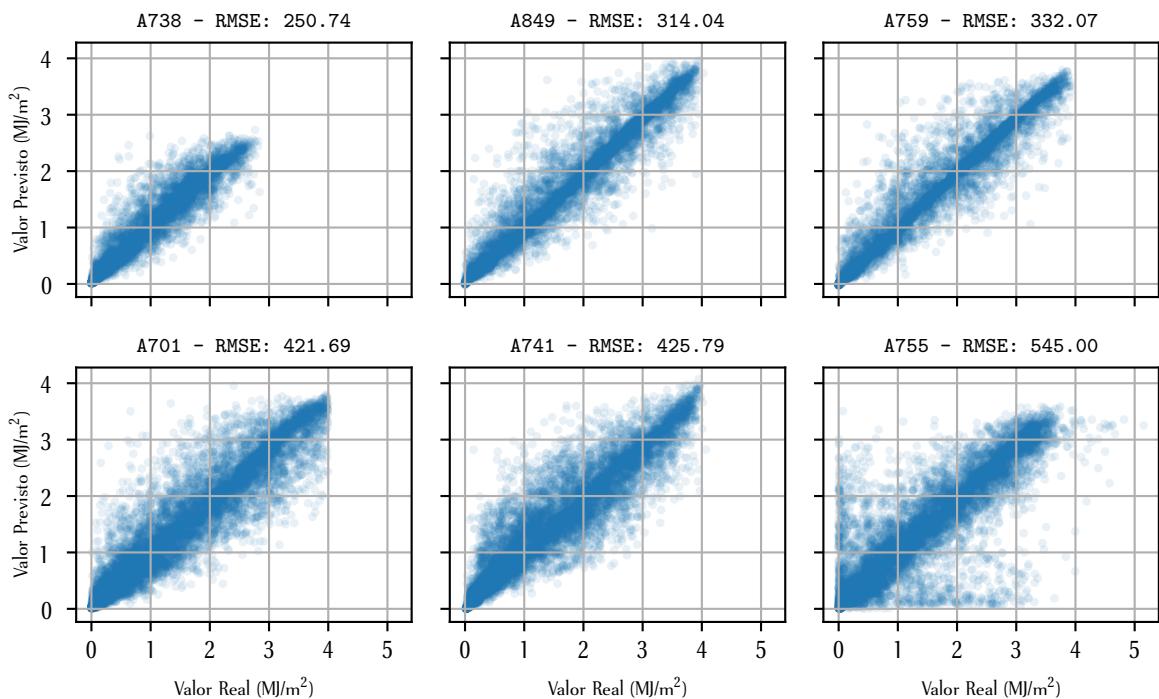
As ilustrações apresentadas a seguir são referentes apenas aos melhores modelos selecionados conforme a Seção 4.4. A Figura 2 sumariza a avaliação de desempenho dos modelos individuais para cada uma das localidades utilizadas para treinamento.

Figura 2 – Desempenho dos modelos em cada localidade utilizando as métricas RMSE e MAE.



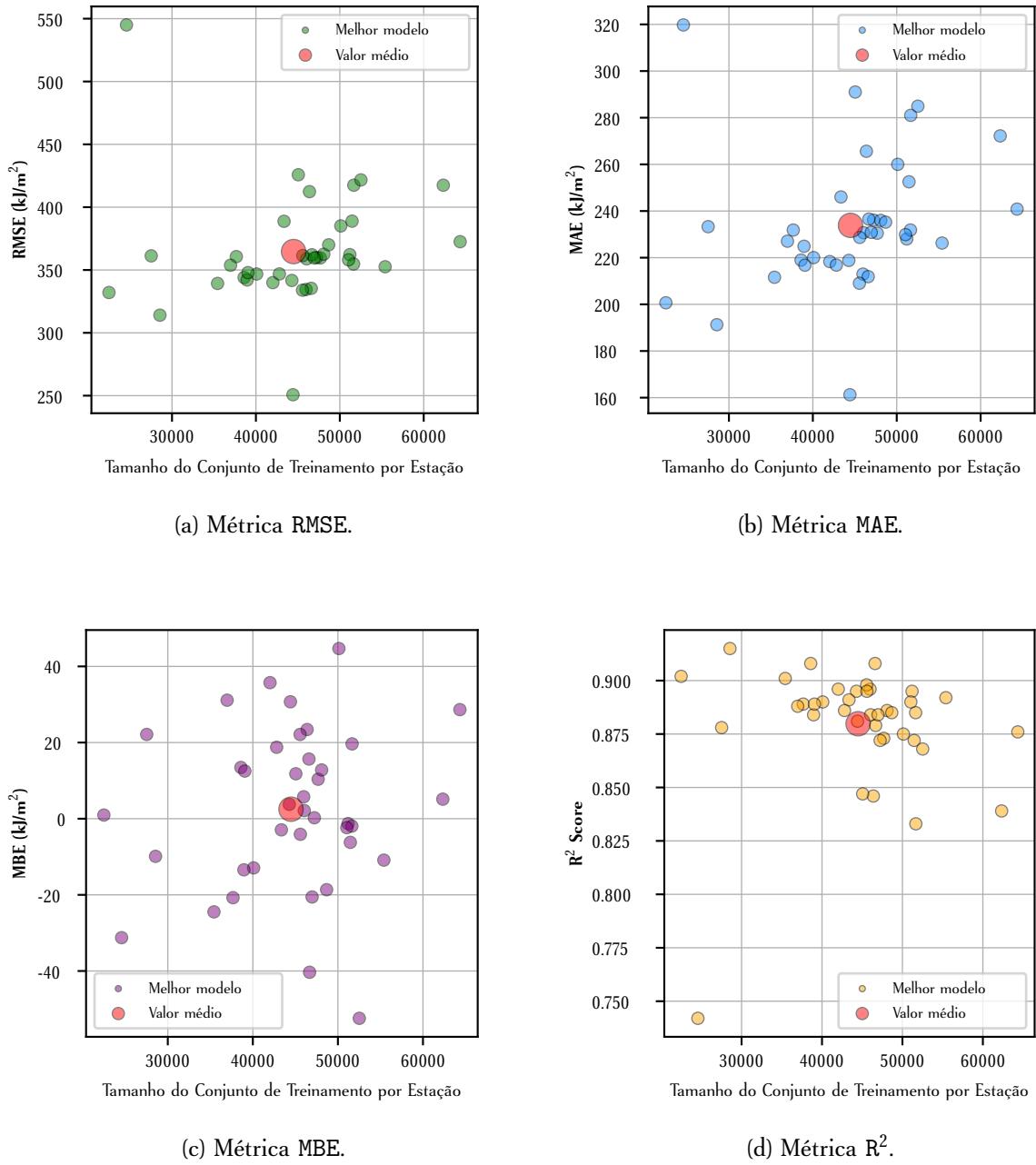
Dentre todos os modelos obtidos durante a seleção, destacaram-se os modelos das localidades apresentadas pela Figura 3 como os modelos com as melhores e piores capacidades de generalização observadas durante o período de avaliação nos conjunto de teste.

Figura 3 – Na parte superior, os três melhores modelos após a avaliação; na parte inferior, os 3 modelos com pior capacidade de generalização, todos em termos de RMSE.



As Figuras 4a a 4d apresentam o desempenho dos melhores modelos em face das métricas de desempenho adotadas em relação à quantidade de observações disponíveis em cada uma dos locais de treinamento.

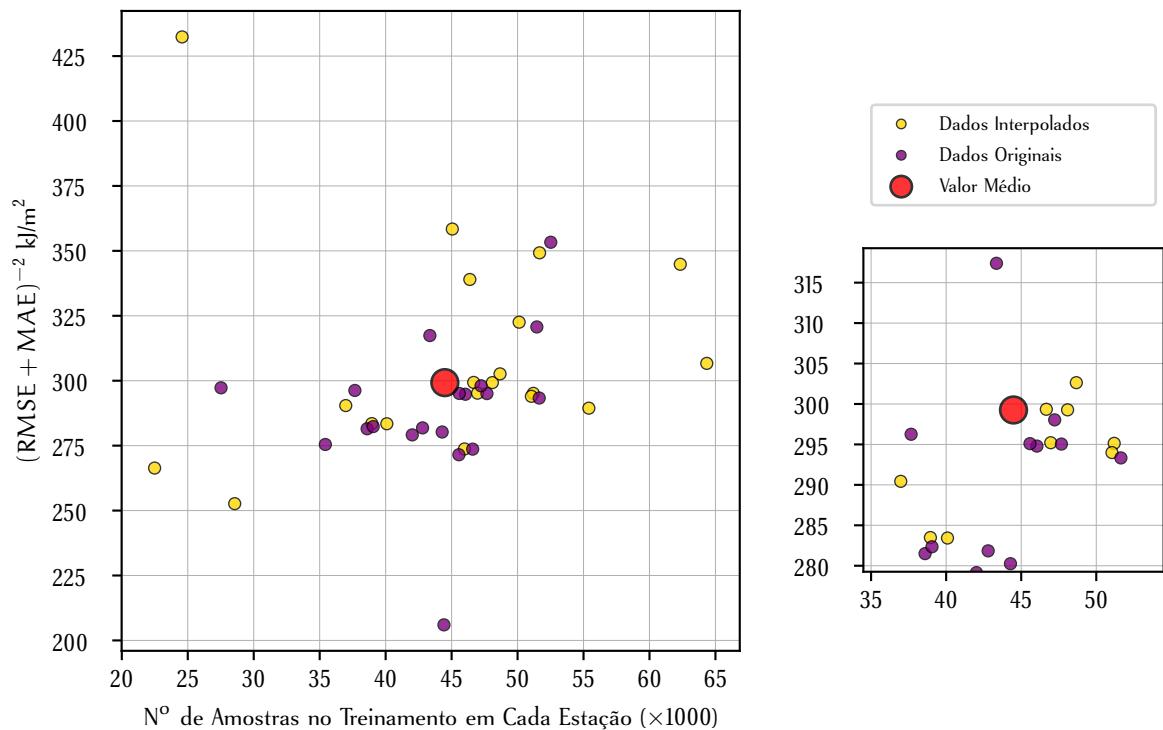
Figura 4 – Métricas de desempenho para os melhores modelos obtidos. Os círculos indicam o melhor modelo de cada local de treinamento.



Fonte: Elaborado pelo autor

A Figura 5 apresenta de forma comparativa as métricas RMSE e MAE para o melhor modelo de cada uma das localidades utilizadas. Para a descrição detalhada do desempenho de cada um dos melhores modelos de cada localidade confira a Tabela 5 do Apêndice B.

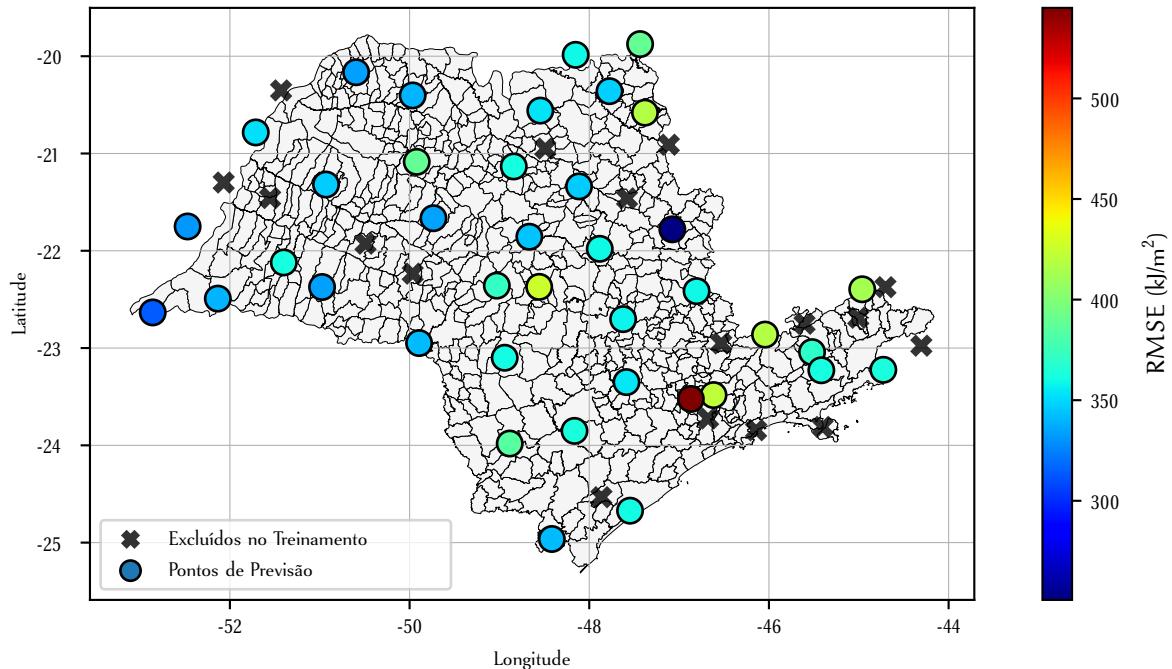
Figura 5 – Média entre as métricas RMSE e MAE para cada modelo. Círculos em amarelo indicam que o melhor desempenho foi obtido utilizando dados com interpolação, enquanto os círculos em roxo indicam que o melhor desempenho foi obtido com os dados originais.



Fonte: Elaborado pelo autor.

A Figura 6 apresenta o desempenho dos modelos em relação à localização geográfica das estações utilizadas no treinamento. Observa-se que o desempenho não é influenciado pela localização geográfica uma vez que não há concentração de modelos com desempenho melhor ou pior em relação ao desempenho médio em regiões específicas. Para a descrição detalhada de cada um dos melhores modelos de cada localidade confira a Tabela 5 do Apêndice B.

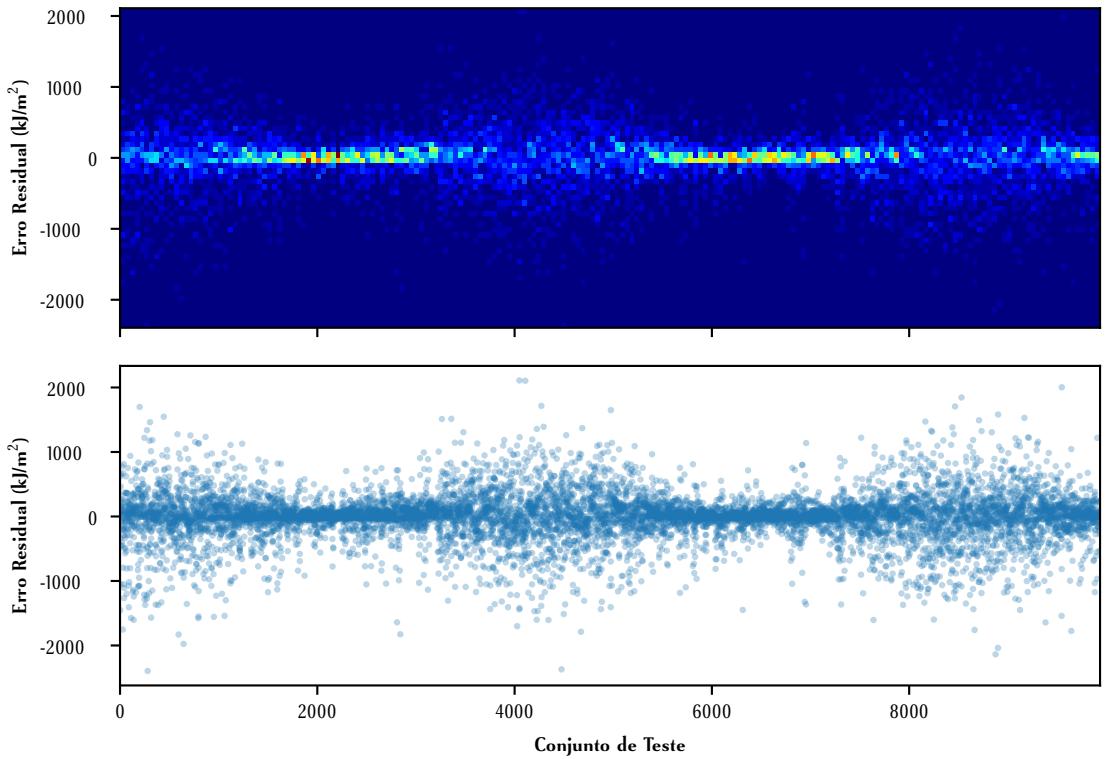
Figura 6 – RMSE em relação à localização geográfica das estações de treinamento.



5.1.1 Plotando Valores Residuais

Durante as etapas de avaliação dos modelos produzidos em cada localidade, foram avaliados os erros residuais nas previsões realizadas durante os testes. Nestas circunstâncias, o conjunto de testes de cada um dos locais utilizados na etapa de avaliação foi ordenado pela data de observação. Esta ordenação foi realizada a fim de verificar a existência de possíveis vieses nos modelos em relação ao período do ano no qual a previsão é realizada

Figura 7 – Padrão presente no erros residuais exemplificado por meio da estação A711.



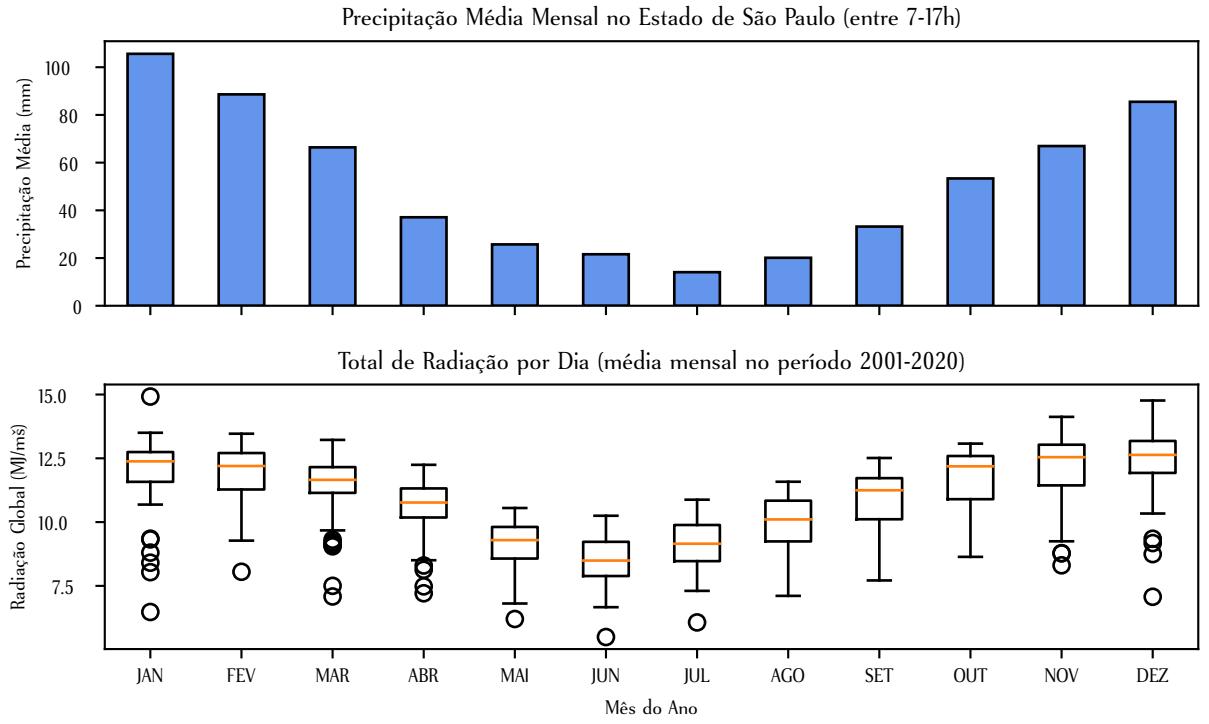
Este padrão, exemplificado por meio da estação A711 na Figura 7, se repete nas demais estações: previsões com maior erro residual próximo às observações múltiplas de 4000 e menores próximas às observações em múltiplos de 2000. Como as observações horárias de cada um dos conjuntos de testes foram ordenadas pela data cada, uma quantidade relativa a um ano em observações seria constituída de:

$$1 \text{ observação por hora} \times 11 \text{ horas por dia} \times 365 \text{ dias no ano} = \mathbf{4015 \text{ observações por ano}}$$

Portanto, o padrão observado está correlacionado com características meteorológicas cíclicas e possui periodicidade de 1 ano. Neste caso, conclui-se que os modelos possuem menores erros residuais em observações entre os meses de Maio e Agosto, já que as observações de todos os conjuntos de testes iniciam-se em Janeiro.

Os maiores erros residuais nos meses de Novembro a Março provavelmente devem-se à maior variação das condições climáticas, sobretudo da precipitação. Estas variações impactam diretamente na capacidade de generalização dos modelos, já que os dados relativos a estes meses são muito mais ruidosos do que aqueles em outros meses do ano. A Figura 8 apresenta a relação entre a quantidade de precipitação no Estado durante o período em que as observações foram selecionadas para o treinamento e a variação na radiação solar.

Figura 8 – Relação entre a precipitação e a variação na radiação solar.



5.2 Avaliação do Modelo de Agregação

A seguir são apresentadas duas abordagens para a extração das previsões dos modelos: uma baseada em uma variação da média ponderada do inverso de distâncias (*Inverse Distance Weighting*) (KESKIN et al., 2015; OZTURK; KILIC, 2016), proposta no Relatório Científico de Progresso (cf. Seção 3.5 deste), e outra utilizando aprendizagem supervisionada com um modelo baseado em *Stacking*. O detalhamento dos resultados obtidos em cada uma das estações utilizadas nos testes de ambos os modelos são apresentados na Tabela 6 do Apêndice C.

5.2.1 Modelo de Agregação com Média Ponderada do Inverso das Distâncias

A avaliação do modelo de agregação proposto foi realizada utilizando-se os conjuntos de dados das estações no período de Janeiro de 2019 até Junho de 2021. Para cada uma das estações, tanto de treinamento quanto de teste, foi obtido um conjunto de dados associando os valores reais esperados para o atributo alvo a uma *timestamp*. Para as estações utilizadas no treinamento (cf. Seção 3.4 do Relatório Científico de Progresso) foram obtidos conjuntos de previsões, realizados pelos melhores regressores em cada local, que seriam comparados com os valores esperados nas respectivas *timestamps*. Em seguida, para cada uma das *timestamps* do conjunto de valores reais esperados foi realizada a agregação das previsões das estações mais próximas, resultando (cf. Seção

3.5 do Relatório Científico de Progresso) em um terceiro conjunto de dados que compara em cada instante de tempo o valor real e o valor previsto pela agregação.

Com relação ao desempenho, este modelo apresentou média de 498,20 kJ/m² no RMSE e média de 0,758 para o R² Score nas estações de teste e produziu previsões com grande amplitude nos valores das métricas avaliadas. Quando comparadas com as médias destas métricas para os modelos específicos em cada local houve um incremento de 36,64% no RMSE e um decremento de 13,8% no R² Score. Para mais informações confira as Tabelas 6 e 5 nos Apêndices C e B, respectivamente.

5.2.2 Modelo de Agregação Alternativo

Como alternativa ao modelo de agregação que utiliza uma variação da média ponderada do inverso de distâncias para realizar a última etapa das previsões, foi utilizado um modelo baseado em *Stacking* que foi treinado utilizando Aprendizagem Supervisionada. Este modelo agregou dois outros: *Extreme Gradient Boosting* e uma Rede Neural — modelos com os melhores desempenhos individuais nos treinamentos dos locais individuais.

Construção do Conjunto de Dados

Para que essa abordagem fosse possível, foi necessário realizar a construção de um novo conjunto de dados a partir dos conjuntos já previamente selecionados para o teste da agregação. Desta forma, a partir do conjunto de teste da agregação original, foram construídos dois novos conjuntos sendo eles: 1 – conjunto de dados de treinamento para o modelo de agregação baseado em aprendizagem supervisionada; e 2 – conjunto de dados para teste deste modelo.

Este conjunto de dados para treinamento foi construído utilizando-se as informações das *timestamps* e das informações dos pontos de treinamento mais próximas. A Equação 1 apresenta o formato de entrada utilizada para treinar o modelo de Aprendizagem Supervisionada; nela D_i, E_i, P_{i_t} e R_t significam a distância entre a i-ésima estação e o ponto de previsão, erro do estimador da i-ésima estação, previsão do estimador da respectiva estação para a *timestamp* t e valor real esperado para aquela *timestamp* t, respectivamente.

$$\underbrace{(\text{DOY}, \text{HORA}}_{\text{timestamp}}, \underbrace{\text{l}_a, \text{l}_o}_{\text{Localização}}, \underbrace{\text{D}_1, \text{E}_1, \text{P}_{1_t}}_{\text{Estação 1}}, \underbrace{\text{D}_2, \text{E}_2, \text{P}_{2_t}}_{\text{Estação 2}}, \dots \underbrace{\text{D}_n, \text{E}_n, \text{P}_{n_t}}_{\text{Estação n}}) \rightarrow \text{R}_t \quad (1)$$

Os limiares de distância mínima e quantidade mínima de estações — 120 km e mínimo de 3 estações, respectivamente — foram mantidos, de forma que o número máximo de estações que satisfaziam a essas condições (n na Equação 1) foi 7. Quando o número de estações mais próximas foi inferior a 7, os valores para distância, erro e previsão das estações faltantes foram preenchidos com 0. Verificou-se que a maioria das amostras continha menos do que 7 estações tornando o

conjunto de dados esparso e introduzindo ruído durante o treinamento. Assim, n foi reduzido para 4 de modo que as amostras que excediam esse número tiveram as estações selecionadas com base na distância.

A separação entre os conjuntos de treinamento e teste foi realizada ordenando os subconjuntos referentes a cada estação de teste da agregação pela *timestamp*, e selecionando iterativamente 4 amostras em ordem para treinamento e 1 amostra para teste. Os subconjuntos de treinamento e de teste para cada estação foram concatenados, resultando nos conjuntos treinamento e teste finais. Essa abordagem foi utilizada para garantir a estratificação dos conjuntos de dados, evitando possíveis vieses que poderiam surgir na seleção aleatória das amostras em relação ao período do ano e a quantidade desproporcional de amostras em cada estação de teste.

Com relação ao desempenho, este modelo apresentou média de 392,33 kJ/m² no RMSE e média de 0,862 para o R² Score. Quando comparadas com as médias destas métricas para os modelos específicos em cada local houve um incremento de 7,58% no RMSE e um decremento de 2,1% no R² Score. Para mais informações confira as Tabelas 6 e 5 nos Apêndices C e B, respectivamente.

5.2.3 Comparação entre as Técnicas de Agregação Utilizadas

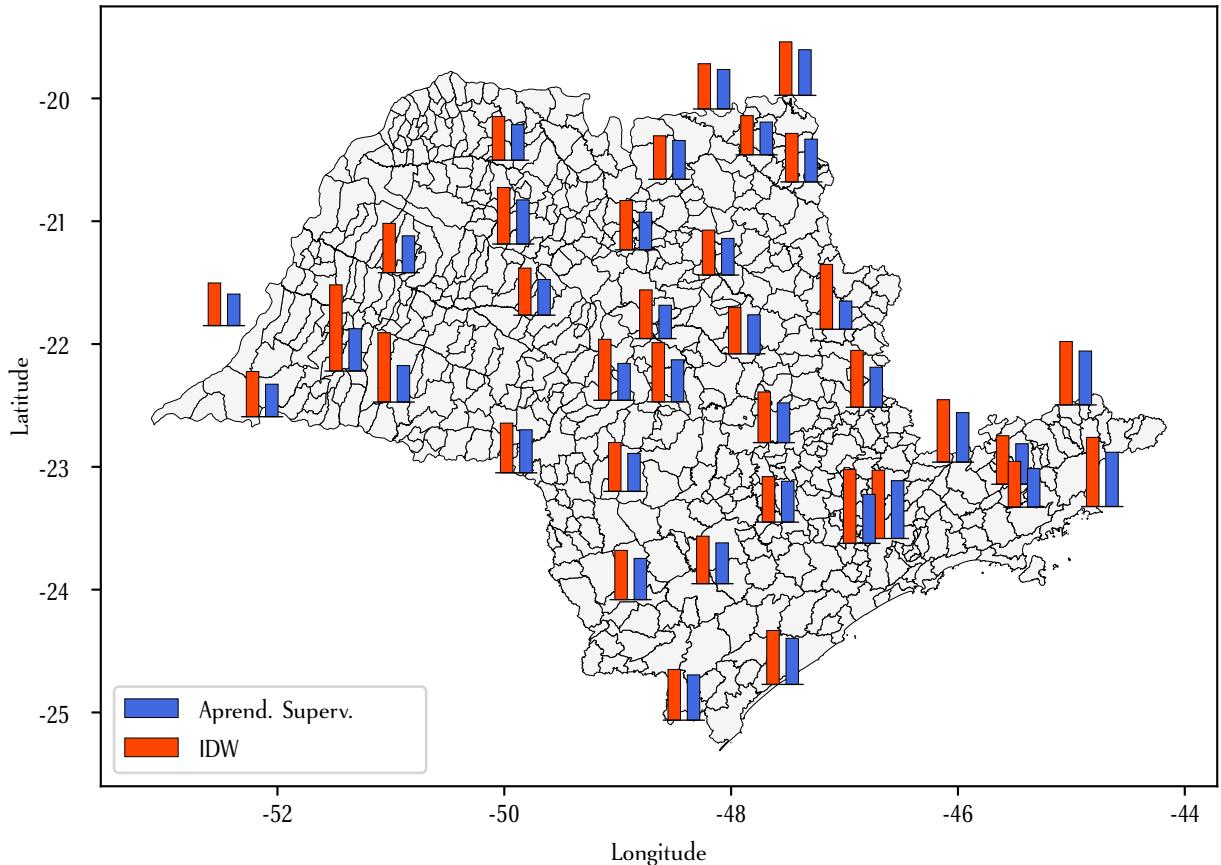
De maneira geral, o desempenho da agregação baseada em aprendizagem supervisionada foi superior à agregação utilizando o inverso das distâncias. A Figura 9 apresenta as comparações do erro médio entre as métricas RMSE e MAE entre os dois modelos, para estações que foram utilizadas no teste de ambas as técnicas em relação à localização geográfica. A Tabela 3 apresenta as diferenças médias para as métricas de desempenho.

Tabela 3 – Comparação entre as métricas de desempenho das técnicas de interpolação utilizadas.

Técnica	RMSE (kJ/m ²)		MAE (kJ/m ²)		MBE (kJ/m ²)		R ² Score	
	\bar{x}	σ	\bar{x}	σ	\bar{x}	σ	\bar{x}	σ
IDW	498,20	107,26	358,73	81,03	-7,55	172,91	0,758	0,130
Apred. Superv.	392,33	57,88	271,29	48,13	25,13	29,34	0,862	0,035
$\Delta\%$	-21,25	-46,04	-24,37	-40,60	432,87	-83,02	13,69	-72,67

Fonte: Elaborado pelo autor.

Figura 9 – Comparação entre os erros para interpolação baseada no inverso da distância (vermelho) e interpolação baseada em aprendizagem supervisionada (azul)



As informações sobre as métricas de desempenho em cada local de teste dos modelos de agregação são apresentados na Tabela 3 do Apêndice C.

6 Conclusões

Ao longo deste projeto foram realizadas diversas etapas de processamento dos dados obtidos: foram construídos *pipelines* tanto para realizar as transformações necessárias no pré-processamento dos conjuntos de dados, quanto para a automatização das rotinas de treinamento e avaliação dos modelos de previsão produzidos.

Em cada local selecionado foram treinados modelos a partir de 5 algoritmos de Aprendizagem Supervisionada diferentes: *Random Forests*, *Extra Trees*, *Extreme Gradient Boosting*, *Support Vector Machine* e Redes Neurais Artificiais. A partir destes algoritmos foram realizadas etapas de busca de hiperparâmetros a fim de melhorar a capacidade de generalização dos modelos.

Os modelos obtidos após o treinamento e hiperparametrização foram comparados entre si e com um modelo de agregação que os reunia (modelo baseado em *Stacking*) a fim de determinar

com base na métrica RMSE o melhor modelo para cada local de treinamento.

Para a extração das previsões realizadas pelos melhores modelos de cada local foram propostas duas abordagens: uma baseada em interpolação espacial e outra baseada em aprendizagem supervisionada por meio da construção de um novo conjunto de dados de treinamento.

O desempenho dos modelos foi avaliada em termos de RMSE, MAE, MBE e R^2 . Os resultados obtidos mostram que a utilização de agregações de modelos estimadores, tanto para estimativas *intra-local* quanto para as generalizações, foi vantajosa por reduzir sensivelmente os valores das métricas de desempenho. Também foi observado que a introdução da interpolação espacial para preencher valores faltantes ou inconsistentes nos dados de treinamento foi proveitosa já que 20 dos 39 locais de treinamento tiveram o desempenho melhorado quando esta técnica foi aplicada.

7 Agradecimentos

Agradecemos ao Departamento de Computação (DComp) da Universidade Federal de São Carlos (UFSCar), *campus* Sorocaba, por prontamente disponibilizar equipamentos que aceleraram a etapa de treinamento dos modelos ao longo do Projeto referente a este Relatório Científico.

Referências

- ACHILLES, R. **Energia Solar Paulista: Levantamento do Potencial.** 2013. P. 8.
- ALZAHRANI, Ahmad et al. Solar Irradiance Forecasting Using Deep Neural Networks. **Procedia Computer Science**, Elsevier, v. 114, p. 304–313, 2017.
- ASSOULINE, Dan; MOHAJERI, Nahid; SCARTEZZINI, Jean-Louis. Quantifying Rooftop Photovoltaic Solar Energy Potential: A Machine Learning Approach. **Solar Energy**, Elsevier, v. 141, p. 278–296, 2017.
- BERGSTRA, James; BENGIO, Yoshua. Random search for hyper-parameter optimization. **Journal of machine learning research**, v. 13, n. 2, 2012.
- EPE, EDPE. Projeção de Demanda de Energia Elétrica-2017-2026. **Ministério de Minas e Energia. Rio de Janeiro**, p. 95, 2017.
- HAYKIN, Simon. **Neural Networks: a Comprehensive Foundation.** Prentice Hall PTR, 1994.
- KESKIN, Merve et al. Comparing spatial interpolation methods for mapping meteorological data in Turkey. In: ENERGY systems and management. Springer, 2015. P. 33–42.
- LI, Jiaming et al. Machine Learning for Solar Irradiance Forecasting of Photovoltaic System. **Renewable energy**, Elsevier, v. 90, p. 542–553, 2016.
- LOU, Siwei et al. Prediction of Diffuse Solar Irradiance Using Machine Learning and Multivariable Regression. **Applied energy**, Elsevier, v. 181, p. 367–374, 2016.
- OZTURK, Derya; KILIC, Fatmagul. Geostatistical approach for spatial interpolation of meteorological data. **Anais da Academia Brasileira de Ciências**, SciELO Brasil, v. 88, p. 2121–2136, 2016.
- PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.
- PERRAULT, Raymond et al. The AI Index 2019 Annual Report. **AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA**, 2019.
- REFAEILZADEH, Payam; TANG, Lei; LIU, Huan. Cross-Validation. In: **Encyclopedia of Database Systems**. Edição: LING LIU e M. TAMER ÖZSU. Boston, MA: Springer US, 2009. P. 532–538. ISBN 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_565. Disponível em: <https://doi.org/10.1007/978-0-387-39940-9_565>.
- SMETS, Arno HM et al. **Solar Energy: The physics and engineering of photovoltaic conversion, technologies and systems.** UIT Cambridge, 2015.

SUYKENS, Johan AK; VANDEWALLE, Joos. Least Squares Support Vector Machine Classifiers.
Neural processing letters, Springer, v. 9, n. 3, p. 293–300, 1999.

WOLPERT, David H. Stacked generalization. **Neural networks**, Elsevier, v. 5, n. 2, p. 241–259, 1992.

ZENG, Jianwu; QIAO, Wei. Short-term Solar Power Prediction Using a Support Vector Machine.
Renewable Energy, Elsevier, v. 52, p. 118–127, 2013.

APÊNDICE A – Estações após a Imputação

A Tabela 4 apresenta os resultados da interpolação em relação à quantidade de amostras disponíveis em cada uma das estações utilizadas como fonte de dados.

Tabela 4 – Métricas de desempenho para o melhor modelo de cada estação.

Estação	Amostras Originais	Amostras com Imputação	Δ	$\Delta\%$
A726	52,818	63,349	10,531	19,94
A509	52,434	69,62	17,186	32,78
A520	57,305	63,466	6,161	10,75
A525	58,96	62,824	3,864	6,55
A529	51,603	60,291	8,688	16,84
A561	22,211	24,601	2,39	10,76
A619	55,84	62,62	6,78	12,14
A628	11,466	15,773	4,307	37,56
A635	2,05	15,696	13,646	665,66
A701	61,663	64,074	2,411	3,91
A704	56,845	64,568	7,723	13,59
A705	70,635	78,669	8,034	11,37
A706	19,628	67,282	47,654	242,79
A707	36,442	49,716	13,274	36,43
A708	66,161	75,655	9,494	14,35
A711	55,657	59,961	4,304	7,73
A712	57,024	58,456	1,432	2,51
A713	60,026	63,584	3,558	5,93
A714	56,766	61,41	4,644	8,18
A715	54,371	63,535	9,164	16,85
A716	51,088	61,515	10,427	20,41
A718	51,965	60,756	8,791	16,92
A725	41,528	63,322	21,794	52,48
A727	54,232	63,014	8,782	16,19
A728	46,995	62,279	15,284	32,52
A729	51,667	61,359	9,692	18,76
A733	53,444	56,824	3,38	6,32
A734	46,962	58,681	11,719	24,95
A735	48,004	57,919	9,915	20,65
A736	45,192	58,455	13,263	29,35
A737	43,173	58,541	15,368	35,6
A738	50,377	59,999	9,622	19,1
A739	50,647	58,617	7,97	15,74
A740	49,998	58,548	8,55	17,1
A741	43,545	56,554	13,009	29,87
A744	12,927	14,372	1,445	11,18
A746	35,904	51,993	16,089	44,81
A747	51,512	56,531	5,019	9,74
A748	41,498	47,166	5,668	13,66
A753	43,991	55,428	11,437	26,0
A755	21,396	43,757	22,361	104,51
A759	26,869	32,423	5,554	20,67
A762	15,813	18,736	2,923	18,48
A763	10,753	16,972	6,219	57,84
A764	13,643	19,376	5,733	42,02
A765	10,893	18,234	7,341	67,39
A766	4,461	17,953	13,492	302,44
A767	9,148	15,042	5,894	64,43
A768	11,545	16,801	5,256	45,53
A769	12,622	15,101	2,479	19,64
A770	7,379	7,647	268,0	3,63
A771	13,018	13,364	346,0	2,66
A849	35,897	35,897	0,0	0,0
A850	44,17	49,38	5,21	11,8
S705	1,617	12,65	11,033	682,31
S717	6,037	12,681	6,644	110,05
Média	37,854	46,661	19,761	56,989

APÊNDICE B - Detalhamento do Desempenho dos Modelos Individuais

A Tabela 5 apresenta as métricas de desempenho para os melhores modelos de cada estação. Ela já contempla a seleção realizada entre os modelos que utilizaram os dados originais e os dados interpolados ambos com os melhores hiperparâmetros encontrados nas etapas de *Grid Search*.

Tabela 5 – Métricas de desempenho para os melhores modelos de cada estação, após o treinamento.

Estação	Algoritmo Utilizado	Métrica de Desempenho				
		RMSE (kJ/m ²)	MAE (kJ/m ²)	MBE (kJ/m ²)	R ² Score	(RMSE + MAE) ⁻²
A509	StackingRegressor	417,478	281,013	19,648	0,833	349,246
A520	StackingRegressor	359,597	230,494	10,367	0,873	295,046
A525	StackingRegressor	388,888	252,522	-6,211	0,872	320,705
A529	StackingRegressor	412,346	265,618	23,415	0,846	338,982
A619	StackingRegressor	362,237	228,034	-1,28	0,895	295,136
A701	StackingRegressor	421,693	284,93	-52,428	0,868	353,312
A704	StackingRegressor	352,585	226,271	-10,882	0,892	289,428
A705	MLPRegressor	372,524	240,827	28,646	0,876	306,676
A707	StackingRegressor	361,24	233,295	22,135	0,878	297,268
A708	StackingRegressor	417,463	272,209	5,13	0,839	344,836
A711	StackingRegressor	358,865	230,721	2,134	0,884	294,793
A712	StackingRegressor	359,983	236,091	0,24	0,872	298,037
A713	StackingRegressor	354,795	231,87	-1,942	0,885	293,332
A714	StackingRegressor	385,079	260,001	44,693	0,875	322,54
A715	StackingRegressor	362,568	235,982	12,802	0,886	299,275
A716	StackingRegressor	341,682	218,826	3,772	0,895	280,254
A718	StackingRegressor	334,468	212,978	5,757	0,896	273,723
A725	StackingRegressor	360,657	231,852	-20,742	0,889	296,254
A726	XGBRegressor	358,103	229,882	-2,347	0,89	293,992
A727	StackingRegressor	335,416	211,888	15,673	0,908	273,652
A728	StackingRegressor	370,028	235,244	-18,647	0,885	302,636
A729	StackingRegressor	339,96	218,334	35,733	0,896	279,147
A733	StackingRegressor	333,948	209,043	22,136	0,898	271,496
A734	StackingRegressor	346,816	220,039	-12,924	0,89	283,428
A735	StackingRegressor	388,757	246,028	-2,927	0,891	317,392
A736	StackingRegressor	362,137	236,549	-40,338	0,879	299,343
A737	StackingRegressor	344,017	218,982	13,46	0,908	281,5
A738	StackingRegressor	250,735	161,254	30,708	0,881	205,995
A739	StackingRegressor	359,555	230,863	-20,558	0,884	295,209
A740	StackingRegressor	361,491	228,705	-4,096	0,895	295,098
A741	StackingRegressor	425,79	291,016	11,802	0,847	358,403
A746	StackingRegressor	342,082	224,895	-13,427	0,884	283,488
A747	StackingRegressor	346,865	216,82	18,765	0,886	281,842
A748	StackingRegressor	353,802	227,071	31,125	0,888	290,436
A753	StackingRegressor	347,917	216,777	12,509	0,889	282,347
A755	ExtraTreesRegressor	544,999	319,757	-31,247	0,742	432,378
A759	StackingRegressor	332,066	200,674	0,956	0,902	266,37
A849	StackingRegressor	314,041	191,278	-9,888	0,915	252,659
A850	StackingRegressor	339,294	211,607	-24,469	0,901	275,45
Média		364,666	233,852	2,494	0,88	299,259

APÊNDICE C – Detalhamento do Desempenho dos Modelos de Agregação

A Tabela 6 apresenta os resultados obtidos com a aplicações dos dois modelos de agregação propostos. Ao todo foram realizados testes em 50 estações para o modelo baseado no média ponderada pelo inverso das distâncias e 36 no modelo baseado em aprendizagem supervisionada. A diferença se deu em decorrência da disponibilidade de observações em cada estação em relação ao período de teste no momento da construção do conjunto de dados de treinamento do modelo baseado em aprendizagem supervisionada.

Tabela 6 – Comparação de desempenho entre o modelo baseado na variação da média ponderada pelo inverso da distância (IDW) e do modelo baseado em aprendizagem supervisionada (AS).

Estação	RMSE (kJ/m ²)		MAE (kJ/m ²)		MBE (kJ/m ²)		R ² Score	
	IDW	AS	IDW	AS	IDW	AS	IDW	AS
A725	454,36	365,03	341,29	253,15	-154,2	-15,42	0,82	0,883
A509	595,14	467,8	421,13	337,56	15,78	38,76	0,648	0,785
A520	441,57	385,4	294,84	257,54	35,61	34,09	0,822	0,864
A525	499,85	436,14	371,71	307,03	-36,4	2,66	0,799	0,848
A529	608,34	512,42	424,49	365,27	-35,2	-5,17	0,717	0,8
A619	662,36	513,33	463,82	371,08	206,07	65,92	0,693	0,816
A701	628,31	537,28	482,96	404,32	-53,7	17,0	0,704	0,783
A705	594,99	355,28	397,9	244,99	316,38	104,89	0,69	0,889
A707	823,54	399,92	579,85	289,2	454,66	109,15	0,384	0,855
A708	468,71	414,89	321,08	281,78	33,43	19,48	0,796	0,841
A711	443,76	381,7	313,11	253,77	-64,87	9,63	0,815	0,863
A712	509,45	434,23	368,65	316,22	73,45	17,84	0,775	0,836
A713	435,55	391,16	305,26	273,72	-9,77	36,14	0,807	0,844
A714	466,35	394,02	335,6	277,39	5,85	19,54	0,831	0,879
A715	443,7	387,4	329,66	277,12	-24,11	-1,14	0,833	0,873
A716	478,32	412,82	330,76	287,4	-50,39	22,6	0,793	0,843
A718	656,75	353,66	472,43	240,21	-302,23	-0,48	0,594	0,883
A726	468,39	382,79	356,31	264,35	-160,15	0,08	0,816	0,877
A727	448,5	348,85	318,24	233,11	-70,69	8,43	0,845	0,906
A728	471,59	390,6	318,48	268,92	18,08	23,46	0,818	0,875
A729	425,01	343,23	286,3	236,74	101,31	15,75	0,835	0,892
A734	464,72	362,56	338,49	237,77	-109,8	24,4	0,806	0,882
A735	543,54	429,0	377,16	289,52	-79,41	12,93	0,79	0,87
A736	463,47	361,75	339,24	251,71	-17,92	-5,03	0,803	0,882
A737	446,31	322,43	346,54	220,34	-211,28	-10,86	0,844	0,919
A738	594,32	273,44	461,93	185,76	442,71	93,03	0,33	0,859
A739	527,82	385,44	398,73	266,83	-265,39	20,12	0,754	0,869
A740	433,26	367,81	311,02	260,97	-82,67	9,63	0,855	0,895
A741	553,14	392,4	411,5	292,5	18,98	20,95	0,722	0,859
A746	493,17	434,42	332,64	302,76	82,71	40,34	0,764	0,818
A747	434,57	356,84	297,2	239,78	-3,91	35,22	0,822	0,88
A748	418,05	378,29	290,89	254,25	-24,62	21,07	0,846	0,875
A753	383,85	323,46	254,41	210,06	3,98	25,46	0,866	0,904
A755	720,42	482,55	483,93	316,18	302,16	56,89	0,523	0,788
A759	400,22	320,01	294,23	194,06	-152,21	27,46	0,848	0,904
A850	422,57	325,65	313,23	203,33	-164,99	10,1	0,848	0,909