UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
DEPARTAMENTO DE COMPUTAÇÃO E TECNOLOGIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
DCT4403 – TÓPICOS AVANÇADOS EM PROGRAMAÇÃO III
PROF. KARLIANE VALE e HULIANE SILVA

TRABALHO PRÁTICO PRIMEIRA UNIDADE

Aluno: Zaú Júlio A. Galvão
Matrícula: 20190022453

GitHub: https://github.com/ZauJulio/LeaATP_III_IA

Dataset: Dados de geração de energia solar: Geração de energia solar e dados do sensor para duas usinas de energia. Os dados foram coletados em duas usinas de energia solar na Índia durante um período de 34 dias. Ele tem dois pares de arquivos - cada par tem um conjunto de dados de geração de energia e um conjunto de dados de leituras do sensor. Os conjuntos de dados de geração de energia são reunidos no nível do inversor - cada inversor tem várias linhas de painéis solares anexados a ele. Os dados do sensor são coletados no nível da planta - uma única série de sensores colocados de maneira ideal na planta.

Descrição do dataset: O conjunto de dados é dividido em 4 tabelas, 2 de dados de geração energética e duas de dados climáticos, ambos coletados no mesmo intervalo temporal. As tabelas são constituídas de séries temporais de 15 dias, com amostragem em intervalos de 15 minutos. Os datasets de geração possuem 22 fontes, correspondentes aos dados coletados pelos sensores anexados aos inversores. Com os seguintes atributos: Data e hora da instância, Corrente contínua gerada, Corrente Alternada gerada, Total diário gerado e Total absoluto gerado, dados coletados pelo inversor. Contendo a seguinte estrutura:

Gen01:
Index: 68778 entries
Data columns (total 5 columns):

#	Column	Non-Null Count	Dtype
0	DATE_TIME	- 68778 non-null	datetime64[ns]
1	DC_POWER	- 68778 non-null	float64
2	AC_POWER	- 68778 non-null	float64
3	DAILY_YIELD	- 68778 non-null	float64
4	TOTAL_YIELD	- 68778 non-null	float64

dtypes: datetime64[ns](1), float64(4)
memory usage: 4.2+ MB

Gen02:
Index: 67698 entries
Data columns (total 5 columns):

#	Column	Non-Null Count	Dtype
0	DATE_TIME	67698 non-null	datetime64[ns]
1	DC_POWER	67698 non-null	float64
2	AC_POWER	67698 non-null	float64
3	DAILY_YIELD	67698 non-null	float64
4	TOTAL_YIELD	67698 non-null	float64

dtypes: datetime64[ns](1), float64(4)
memory usage: 4.1+ MB

Os datasets de dados climáticos contém o mesmo intervalo de 34 dias, com amostragem em intervalos de 15 minutos, no mesmo período dos datasets de geração de energia. Com os seguintes atributos: Data e Hora da instância, Temperatura ambiente, Temperatura do Módulo e Irradiação. Contendo a seguinte estrutura:

Weather01:
Range Index: 3182 entries
Data columns (total 4 columns):

#	Column	Non-Null Count	Dtype
0	DATE_TIME	3182 non-null	datetime64[ns]
1	AMBIENT_TEMPERATURE	3182 non-null	float64
2	MODULE_TEMPERATURE	3182 non-null	float64
3	IRRADIATION	3182 non-null	float64

dtypes: datetime64[ns](1), float64(3)
memory usage: 149.3+ KB

Weather02:
Range Index: 3259 entries
Data columns (total 4 columns):

#	Column	Non-Null Count	Dtype
0	DATE_TIME	3259 non-null	datetime64[ns]
1	AMBIENT_TEMPERATURE	3259 non-null	float64
2	MODULE_TEMPERATURE	3259 non-null	float64
3	IRRADIATION	3259 non-null	float64

dtypes: datetime64[ns](1), float64(3)
memory usage: 152.9+ KB

a. Descreve o problema da base de dados escolhida (o objetivo).

Perfilar a geração de energia diária com base nos dados históricos.

Correlacionar os dados climáticos com o histórico de geração.

Prever a geração de energia com base em dados climáticos fictícios.

Identificar problemas ou baixo desempenho de um painel.

b. Quantos atributos e quantas instâncias possui a sua base de dados?

Dataset dividido em 4 tabelas, 2 de geração elétrica e 2 de dados climáticos:

Gen01:
Index: 68778 instâncias
Atributos: 5

Gen02:
Index: 67698 instâncias
Atributos: 5

Weather01:
Index: 3182 instâncias
Atributos: 4

Weather02:
Index: 3259 instâncias
Atributos: 4

c. Descreva a caracterização dos dados (tipo, escala)?

Cada dataset possui um atributo temporal de quando cada conjunto de instância foi coletada, em data, hora e minuto, e os demais atributos em ponto flutuante. Os dados de geração estão na escala dos milhares, enquanto que os dados climáticos estão na grandeza das dezenas e uma unidade. O que não deve ser um problema, pois serão estudados separadamente durante a perfilação.

d. A base de dados escolhida contém atributos faltosos? Se sim, quantos?

Sim, inferior a 200 instâncias ausentes.

Qual o pré-processamento realizado diante deste problema?

Interpolação temporal nas colunas e linear nas linhas dos datasets.
https://github.com/ZauJulio/LeaATP_III_IA

e. É necessário normalizar a base de dados? Justifique sua resposta.

Não, ocorrerá perfilação e uso dos perfis para aplicação dos algoritmos.

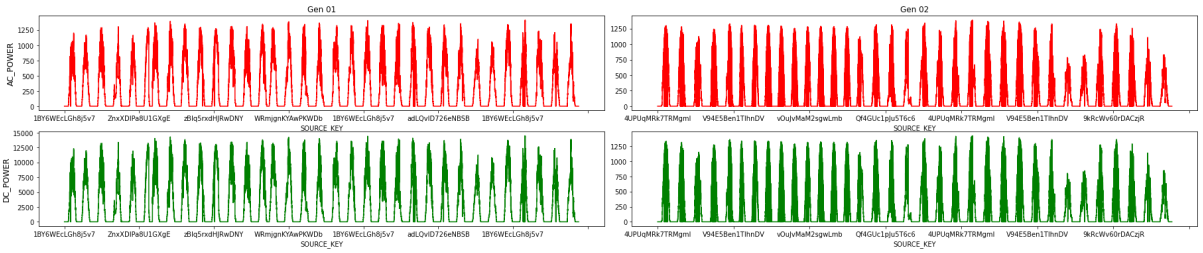
f. É necessário realizar a eliminação manual de atributos, integração de dados, amostragem de dados, balanceamento de dados, limpeza de dados, redução de dimensionalidade e transformação de dados? Descreva cada técnica utilizada.

Sim, dados como 'PLANT_ID' e 'SOURCE_KEY', assumem os mesmos valores para todas as instâncias, respectivamente em geração e dados climáticos. O dataset gen02 foi pré-processado de maneira incorreta, adicionando valores válidos para substituir os faltosos, logo, será completamente descartado.

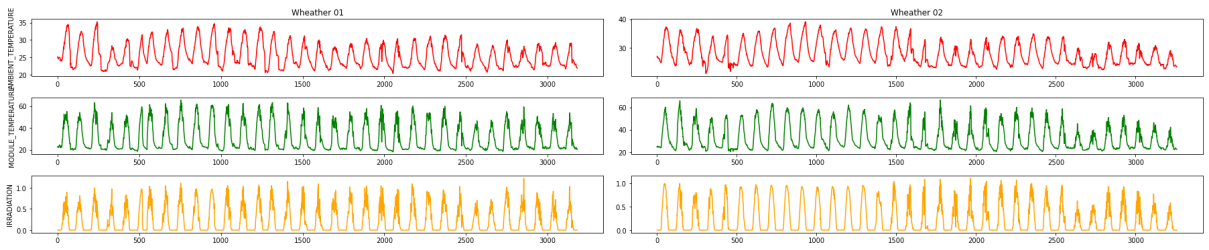
Os dados possuem natureza 3D, contudo simplesmente dividindo-os por fonte dos dados (inversor, 'SOURCE_KEY'), é possível contornar este problema. Como a escala dos dados é indiferente, não será necessário balancear ou transformar os dados. Os conjuntos apresentam poucas instâncias, não será necessário realizar a amostragem.

2- É possível utilizar alguma técnica de visualização de dados antes da etapa de pré-processamento de dados? Se sim, qual/quais? Justifique sua resposta.

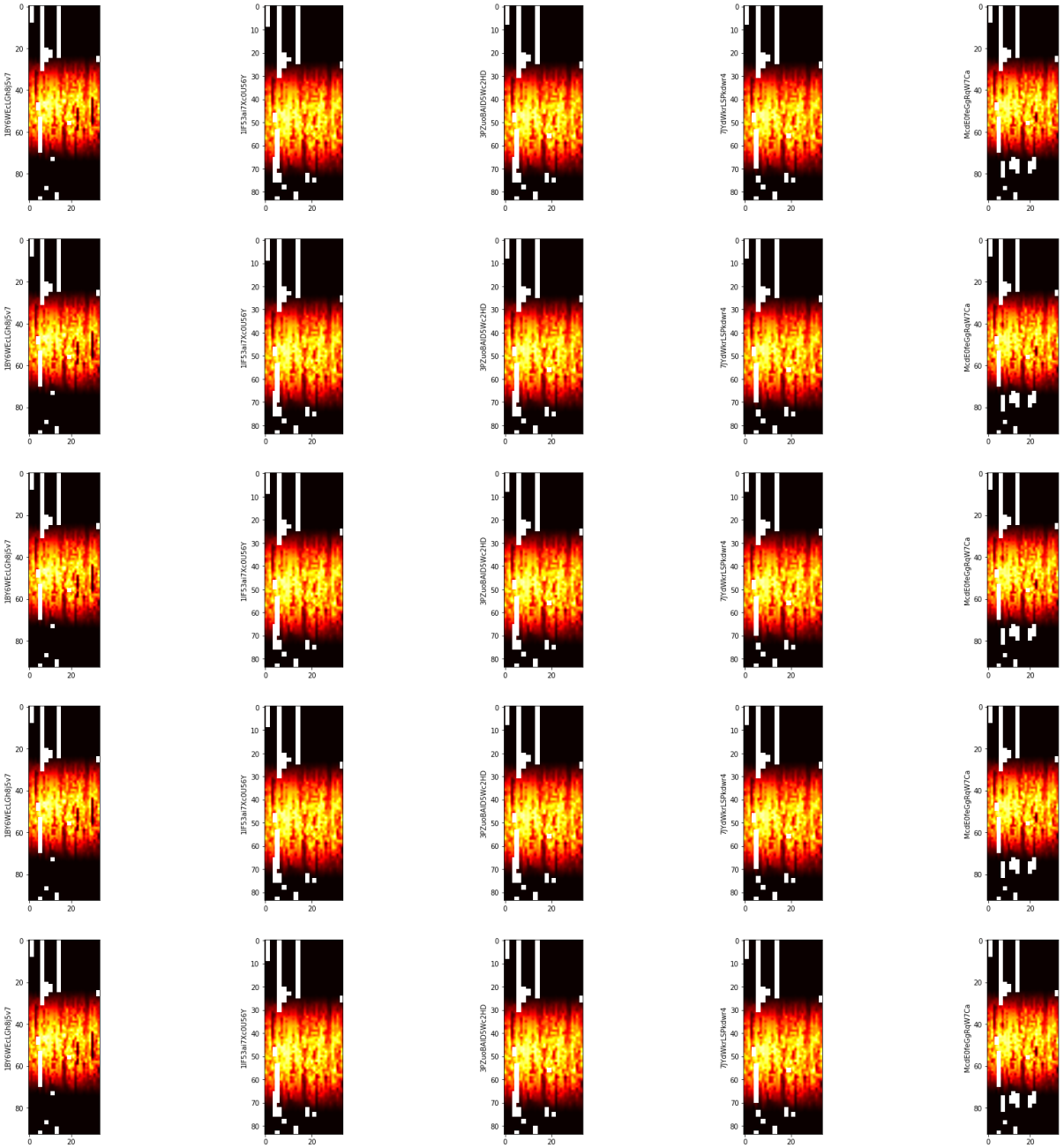
Visualização inicial dos dados de geração energética (Gen01, coluna 1. Gen02, coluna 02. Corrente contínua, linha 1. Corrente alternada, linha 2). Eixo Y corresponde a potências, eixo X ao intervalo temporal.



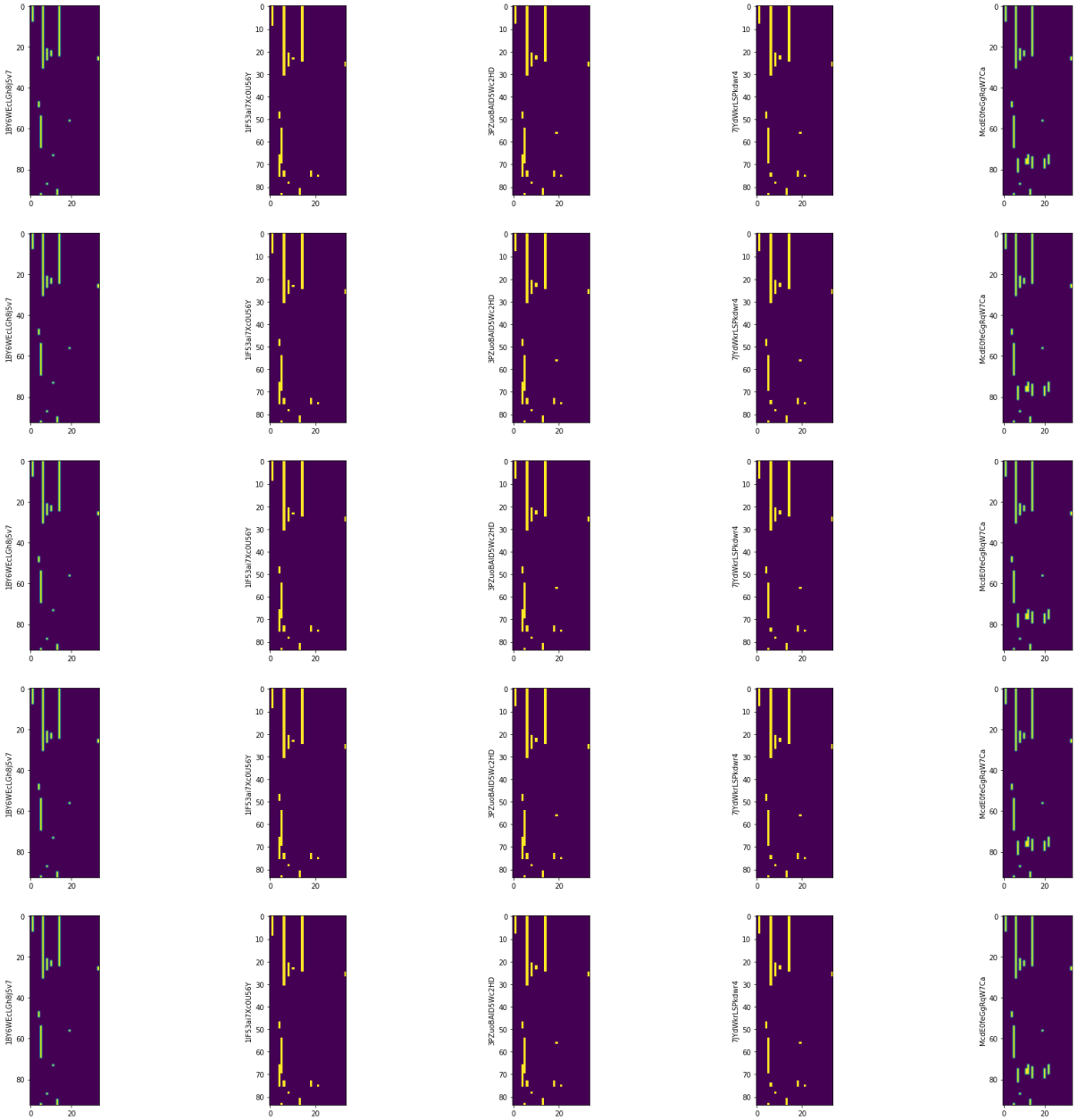
Visualização inicial dos dados climáticos (Weather01, coluna 1.Weather02, coluna 02. Temperatura ambiente, linha 1. Temperatura do módulo, linha 2. Irradiação, linha 3).



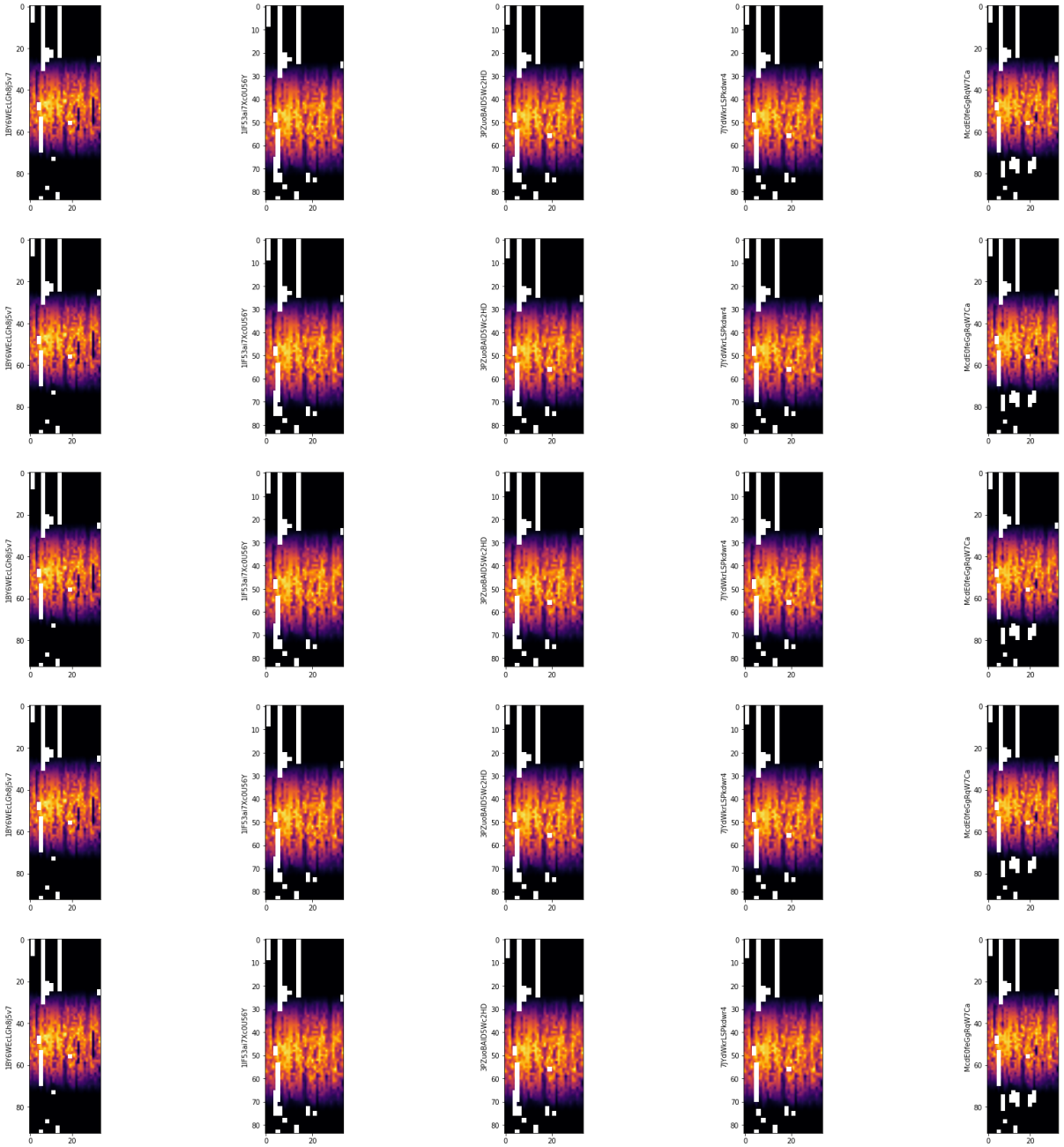
Gen01, Corrente alternada, agrupados por inversor ('SOURCE KEY'). No eixo Y o intervalo, X os dias de amostragem, codificado nas cores a potência AC. Em branco é possível perceber os dados faltosos.



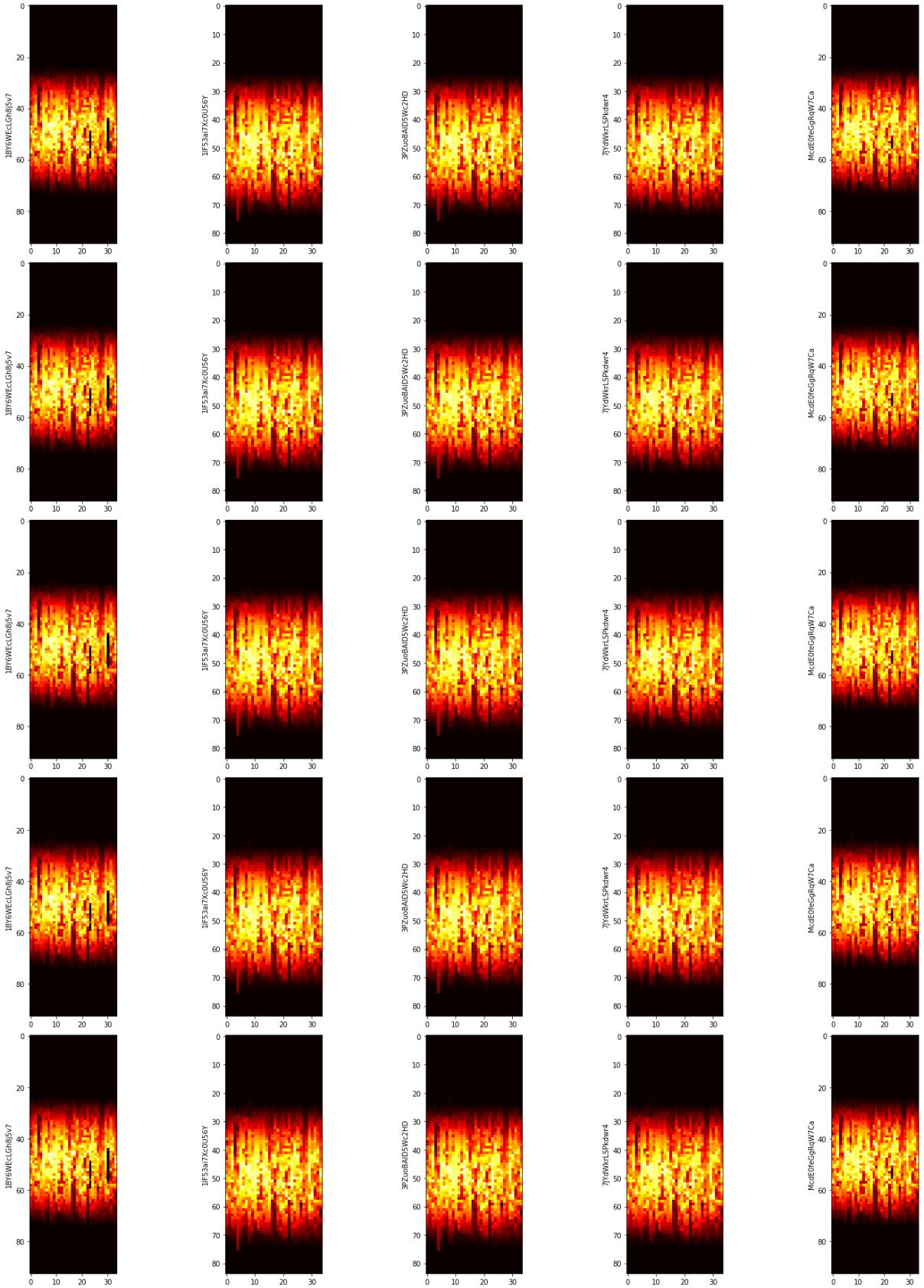
Uma versão do gráfico anterior dedica a visualização somente dos dados faltosos.



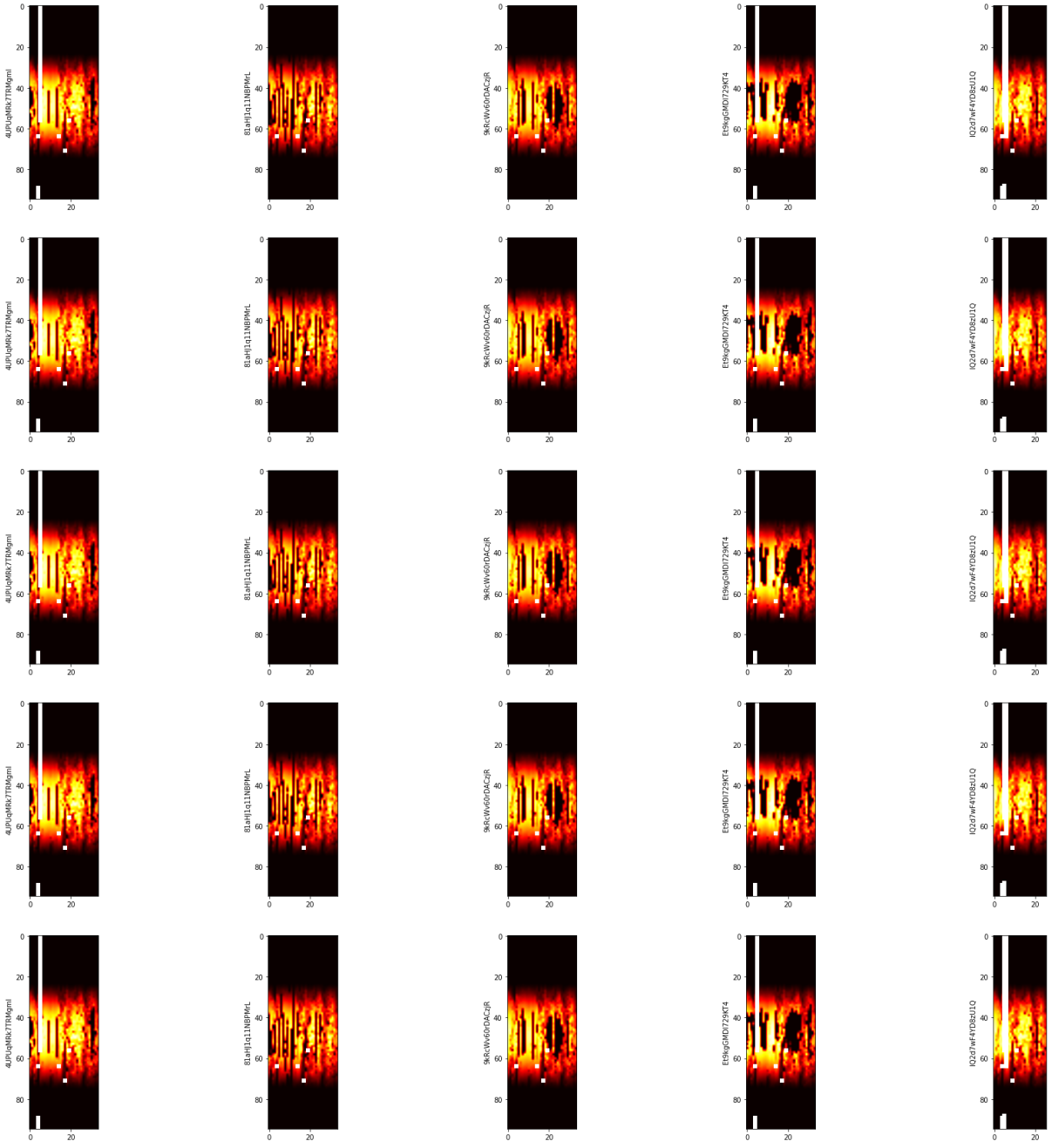
Mesma configuração dos gráficos anteriores. Gen01, corrente contínua.



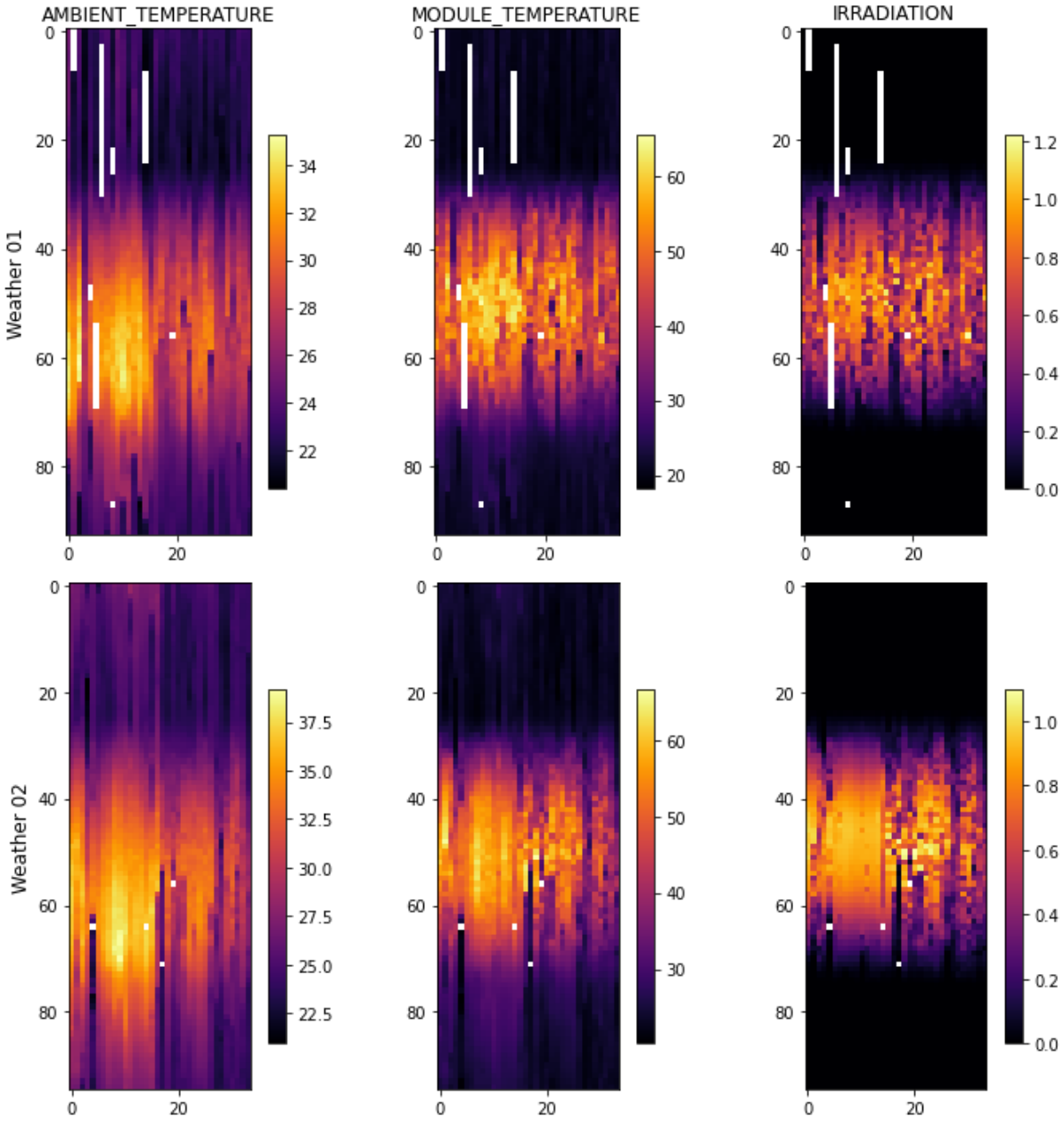
Resolução dos problemas com dados faltosos, utilizando interpolação temporal nas colunas (por dia) e linear nas linhas (no intervalo total) .



Gen02, corrente alternada. É perceptível que houve pré-processamento em vários intervalos durante o dia (pontos pretos, zeros), o problema ocorre também com os outros atributos. Por este motivo o GenO2 foi descartado.



Weather01 linha 1, Weather02 linha2. Dados climáticos, temperatura ambiente, do módulo e irradiação, respectivamente, da esquerda para a direita. Temperatura em graus Celsius. É notável os dados faltosos em branco.



Weather01 linha 1, Weather02 linha2. Dados climáticos, temperatura ambiente, do módulo e irradiação, respectivamente, da esquerda para a direita. Temperatura em graus Celsius. Resultado pós interpolação com a mesma técnica utilizada nos dados de geração energética.