



Presentation Team Imputation

BY: RAMON, MICHAEL, JESÚS, ALBERT, ADRIEN & JULIËN

Table of contents

- Pipeline
- Imputation Methods
- Neural Networks
- Research paper

Pipeline

- Uniform workflow
- File statistics
- Gap creation
- Multiple imputation methods
- Evaluation

The screenshot displays the JupyterLab Pipeline interface. At the top, the header shows the JupyterHub logo, the name 'pipeline', and the status 'Last Checkpoint: 10/26/2021 (unsaved changes)'. On the right, there are buttons for 'Logout' and 'Control Panel'. Below the header is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. A toolbar contains icons for file operations, a 'Run' button, and a 'Validate' button. The main workspace is divided into three sections:

Pipeline

This pipeline is intended to simplify the whole process of loading a dataset, creating gaps of different types in it, imputing the missing data and evaluating the imputation method. You may want to edit the cells preceded by an **EDIT**: sign to fit your needs.

Install dependencies

```
In [1]: !pip install openpyxl
!pip install jupyterlab-widgets
!pip install jsfileupload
!pip install pyxlsb
```

Set an arbitrary random state

```
In [1]: import random

# TODO: this doesn't seem to work properly for functions other than random()
random.seed(7094406398089273)
```

Load clean data

The upload form only supports files up to 10Mo. For larger files, please upload them directly to JupyterHub and provide a relative link to them in the text input herebelow.

EDIT: customize the date parser function.

```
In [2]: from datetime import datetime

def custom_knmi_date_parser(hours: str) -> str:
    return datetime.fromtimestamp(1554109200) + timedelta(hours=int(hours))

def knmi_date_parser() -> str:
    # TODO
    pass
```

Evaluation methods

- Bias
- Sum of error
- Distribution of error
- Maximum error

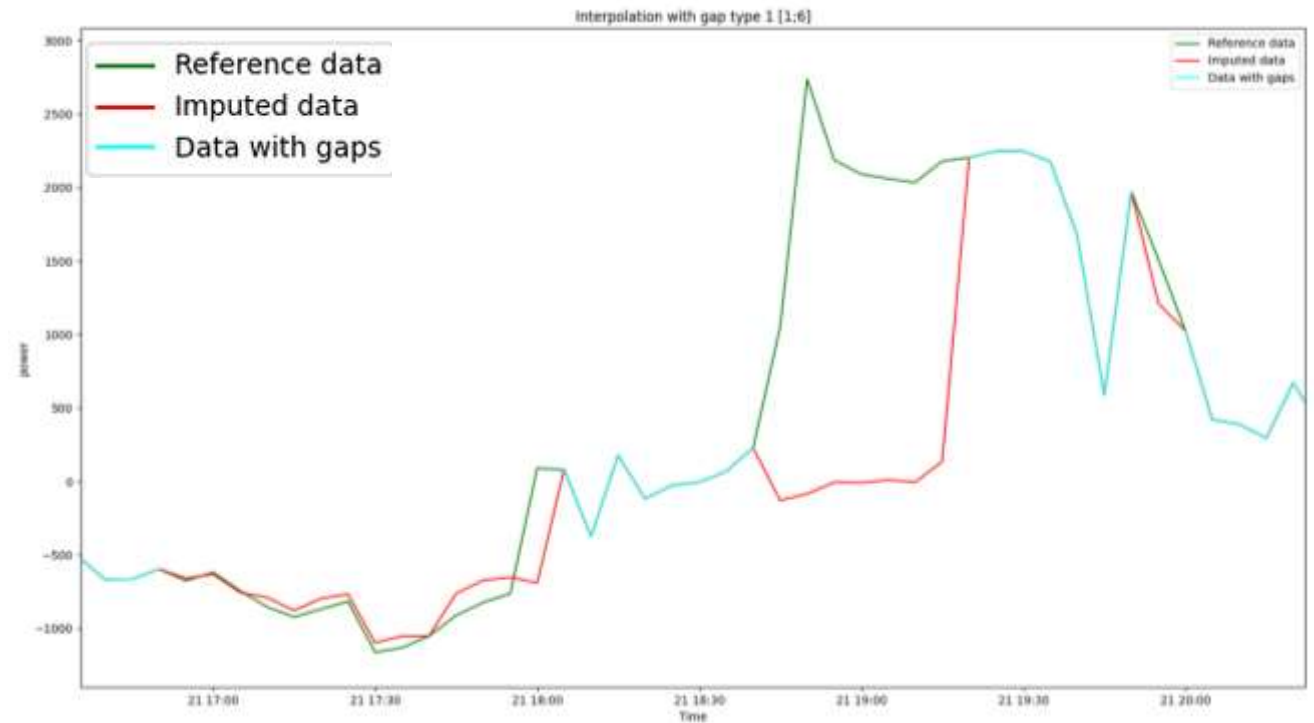
- Kurtosis : measures "tailedness"
- Skewness: measures asymmetry in probability distribution

```
Error distribution with gap type 1 [1;6]
  Mean Squared Error (lower is better)
    357117.4727332921
  Raw Bias
    -108.59066446031686
  Absolute Raw Bias (lower is better)
    628.6086817371577
  Percent Bias (bellow 5% is acceptable)
    15.819181324715185
  Sum (lower is better)
    12225181.642424243
  Maximum (lower is better)
    6379.9000000000001
  Variance (lower is better)
    954858.1722140332
```

Pipeline evaluation metrics for Hot Deck with random donor pool.

Imputation methods

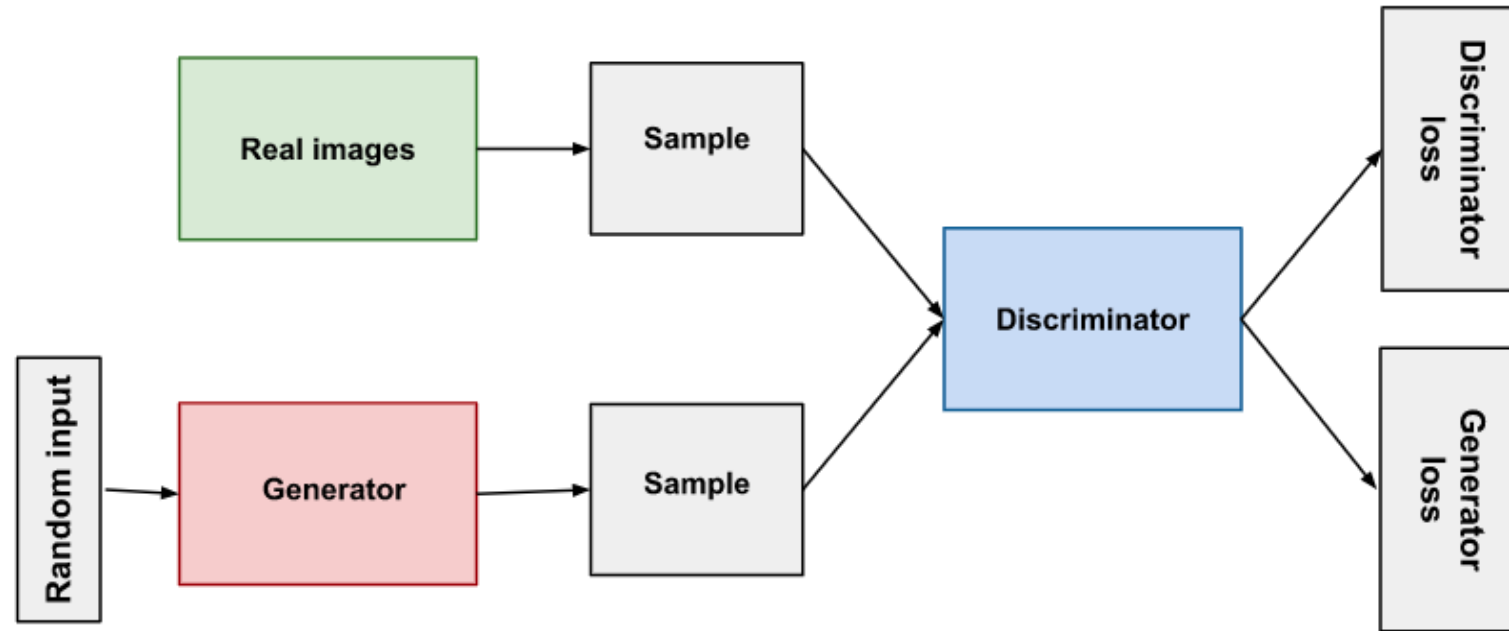
- Mode , Median & Mean
- KNN Nearest Neighbour
- Hot deck
 - Donor pool selection



Hot deck imputation result with a random donor

Neural Networks

- Recurrent Neural Networks
- GAN
- Convolutional Neural Networks



Research (paper)

- Worked on introduction for official paper
- CLIMA conference

What we will be working on

- Evaluation metrics in pipeline
 - Other imputation methods
 - Neural networks
 - Research paper
-
- Focus right now: data collection

Questions & Feedback

- Questions we have for you: