

Contributors

Jack Arnold • Natalie McGuckin • Zuhair Ahmed • Rashmi Ananth • Narmadha Balraj
Skanda Bharadwaj • James Hubbard • Mohammad Issaq • Tucker Kelly • Andy Ko
Sandra Lee • Kyle Liu • Justin Mathew • Elias Winters • Jess Zhang

Project Description

CAT Digital

- Caterpillar designs, manufactures, and sells construction and mining equipment
- The digital branch brings advanced analytics and AI capabilities to the famous yellow iron

CAT 797F Monitoring Service

- Over 70 channels of data sampled each second
- Common problem: missing critical data for a time period due to sensor glitches/anomalies

Data Imputation

- We aim to impute (assign values) to this missing data
- An analytics model in Python will be used to solve this problem



Our data is from a CAT 797F mining truck

Project Purpose

- CAT provides machine condition monitoring services to aid dealers and customers
- Missing gaps of data are problematic for machine learning (ML) models
- Improved ML models will make the monitoring service more reliable

About The Data

Multivariate Time Series Data

- 78 channel sensors per asset
- Sampled Every Second (1Hz frequency)
- Roughly 20 million lines of data per asset

Obfuscated Data

- All channels and assets are given generic names and units to keep data secure
- Ex: Asset ABC00123, Sensor1A (Units C)

MICE Model

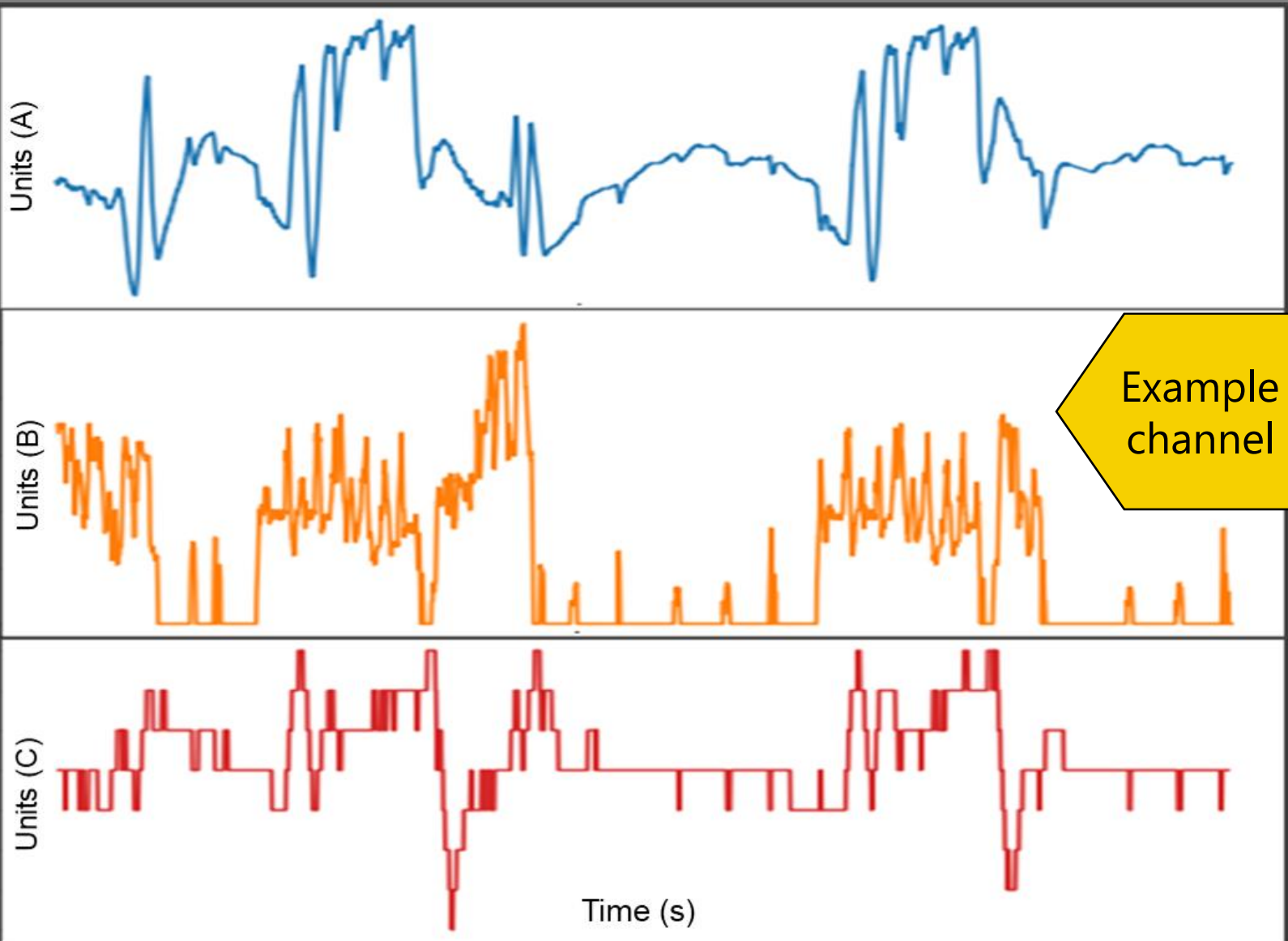
Multiple Imputation by Chained Equations

Iterative Process

- Initial prediction
- Regress on other variables to improve imputed value
- Repeat until convergence

Reasoning

- Works very well when other correlated channels are available
- Accounts for uncertainty due to missing data
- Used for most imputation when similar channels were available



Example 1 Hz channel data

KNN Model

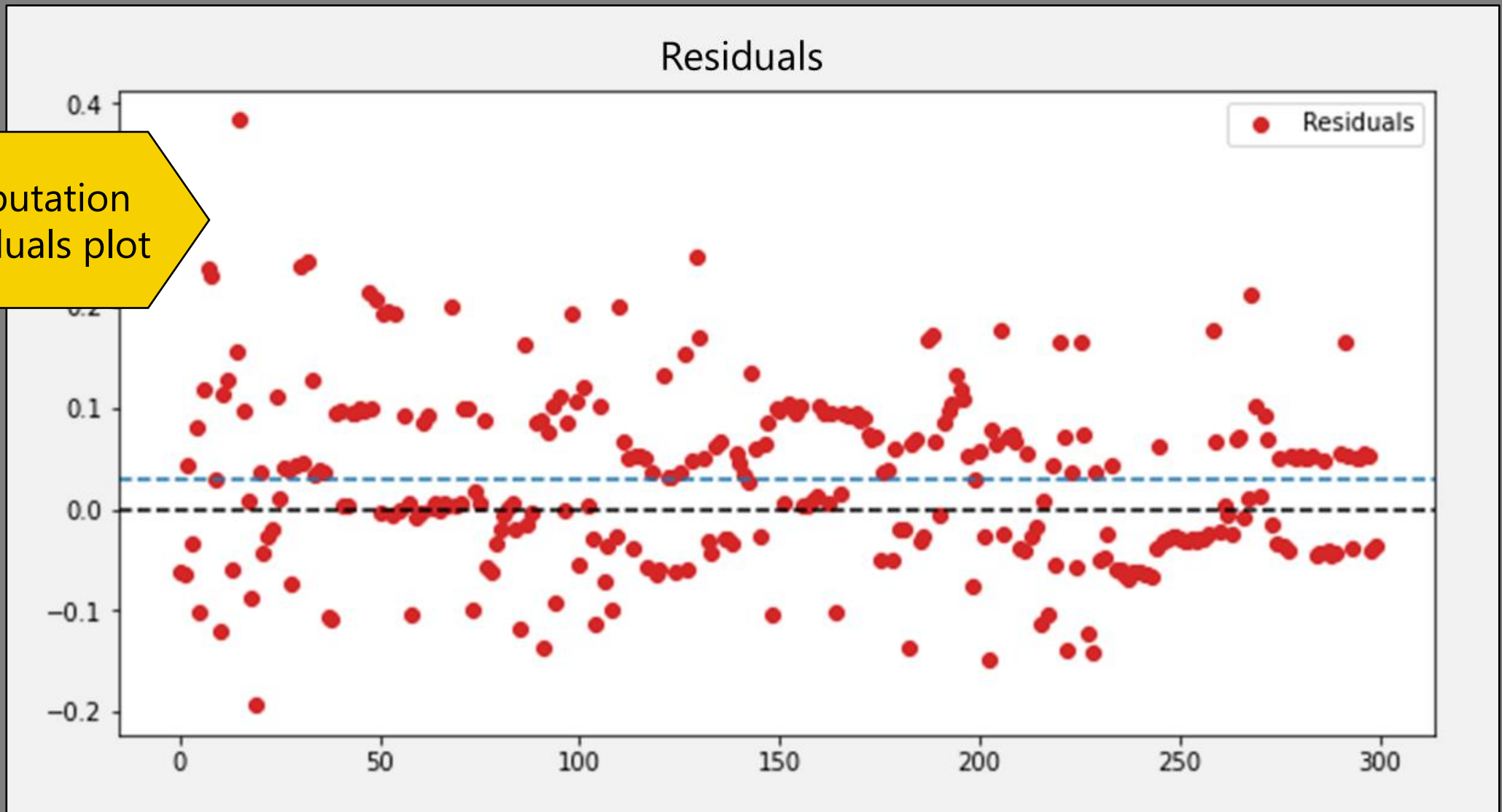
K Nearest Neighbors

- Impute values based on Euclidean distance with other channels
- Projects values into high dimensional space
- This is an effective but memory hungry algorithm

Results & Conclusions

- MICE is a more optimal solution when considering its speed and flexibility
- MICE outperforms traditional imputation methods for over 80% of channels
- On average, MICE has a Mean Absolute Error of less than 10%
- This will solve a significant percentage of the problem

Imputation residuals plot



Cross Validation

Cross Validation

- Simulate missing data using existing data
- Used to compare error statistics across different models

Error Statistics

- Mean Absolute Error
- Bias

Baseline Comparison

- How much of an improvement is the model compared to... ?
 - Mean/Median/Mode Fill
 - Linear Interpolation
- Allows us to determine if our complex approach is worthwhile



The 797F weighs 1.2 million lbs.

Future Goals

Python API

- Callable functions to impute data
- Select the best performing model based on the data characteristics
- Easy to use for CAT Data Scientists

Improve Efficiency

- Formal assessment of tradeoff between time and accuracy
- Allow different levels of precision to be specified in the API
- Improved recognition for channels that cannot be imputed well using our models