Imputing missing data from stock time series

Khanh Nguyen, Yizhen Zhao,

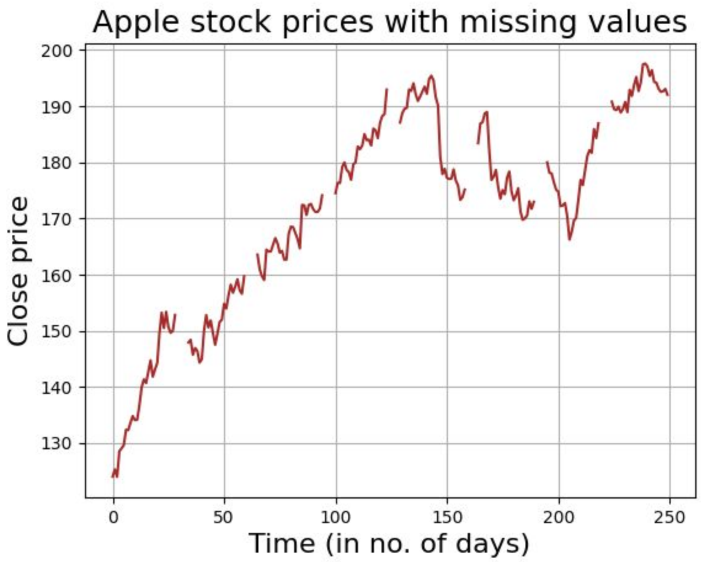
Evgeniya Lagoda, Himanshu Raj,

Carlos Owusu-Ansah, Sergei Neznanov

THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they love at every stage of their career.

Goal: Find the best imputing models for AAPL stock price in 2023

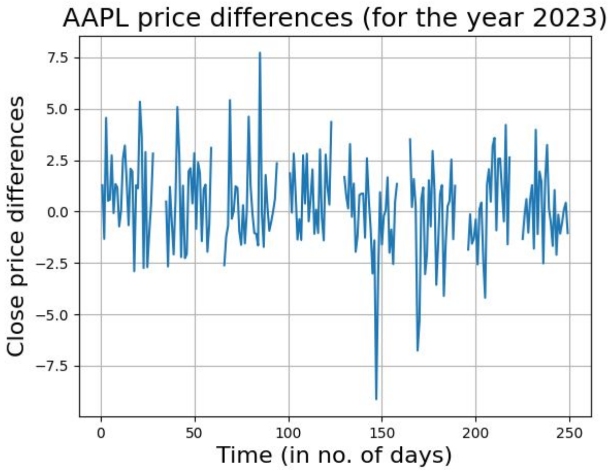


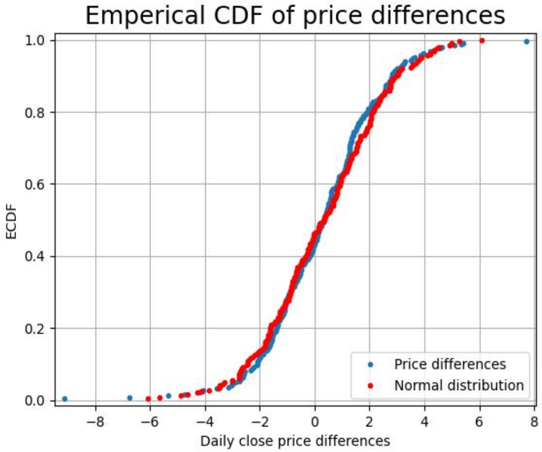
In this plot we removed 7 intervals of 5 consecutive data points

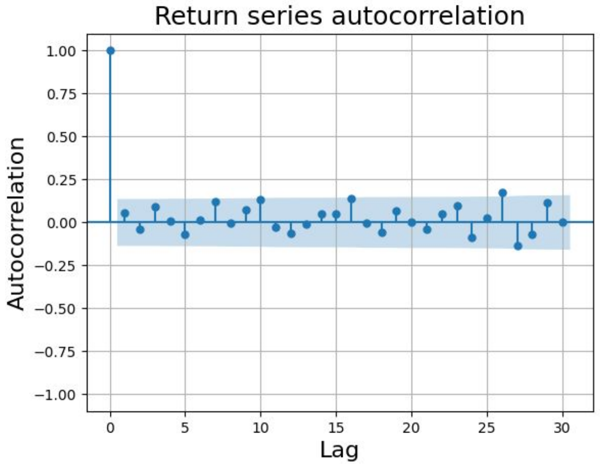
We consider 1, 2, 3, 4, and 5 missing consecutive days.

Date	Open	High	Low	Close	Volume
2023-01-03 00:00:00-05:00	129.215470	129.830399	123.155395	124.048042	112117500
2023-01-04 00:00:00-05:00	125.853183	127.608724	124.057975	125.327515	89113600
2023-01-05 00:00:00-05:00	126.091211	126.725981	123.740581	123.998451	80962700
2023-01-06 00:00:00-05:00	124.980372	129.225391	123.869520	128.560867	87754700
2023-01-09 00:00:00-05:00	129.403910	132.319889	128.828647	129.086517	70790800

Exploratory data analysis







<u>Summary statistics:</u>

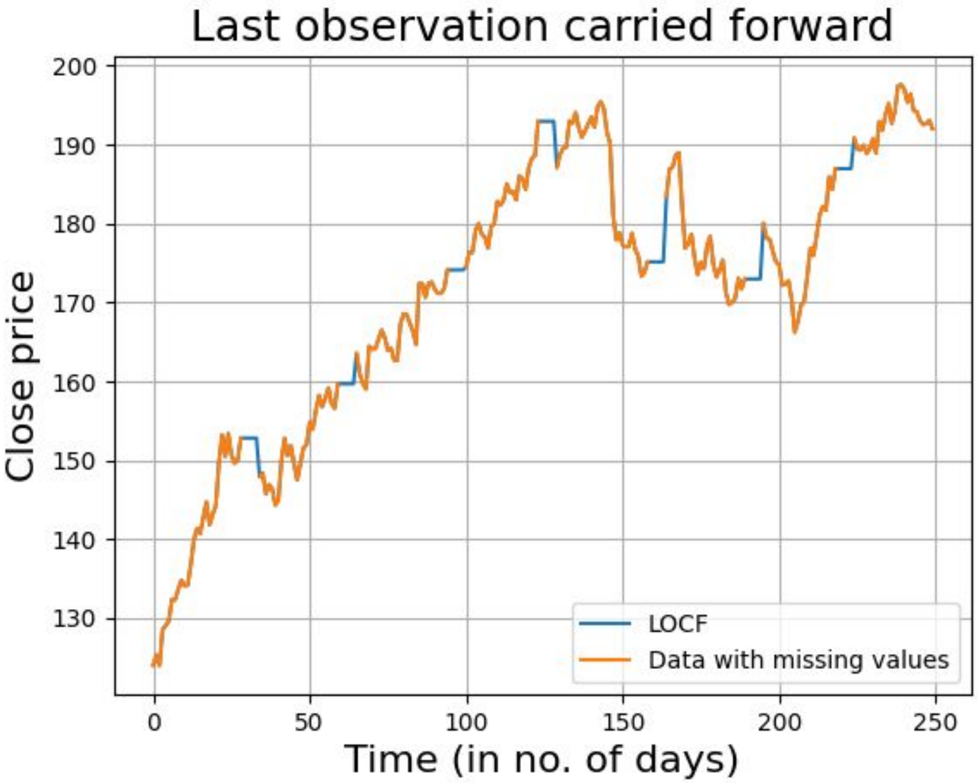
Mean	0.272
Standard deviation	2.11
Skewness	-0.257
Excess kurtosis	1.82

The prices follow a **random walk** where the price differences are **approximately normally distributed** of a large range of price differences.

Since the excess kurtosis is positive (i.e., there are outliers), these differences actually follow a **fat-tailed** distribution

This is reflective of the **Efficient Market Hypothesis** where the simplest possibility is to predict the last value: **Last Observation Carried Forward**

Last observation carried forward



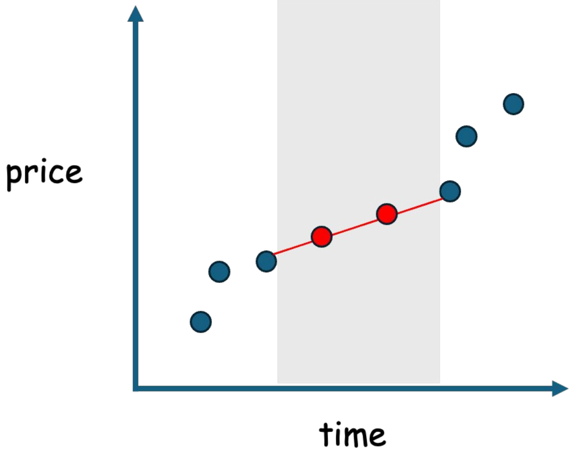
Comparing to the true data this yields an MSE of 11.138

Can we do better?

Models

Models	**Apple's Features**	**Other features**
Linear Interpolation	*Close values*	*None*
Rolling Average	Close values	None
Double Exponential Smoothing	Close values	None
SARIMA / Average over forward & reverse ARIMA	Close values	None
KNN(2)	Open values and dates	None
Regression	Close Daily Return, Close	Close Daily Return NVDA, MSFT, TSM, META, GOOG
VAR(9)	Close Difference	NVIDIA's Close Difference

Baseline – Linear Interpolation



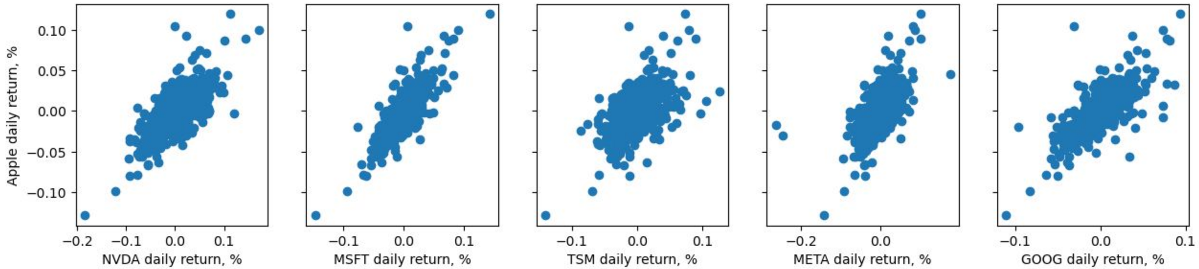
MSE: 3.475

Certainly better than LOCF

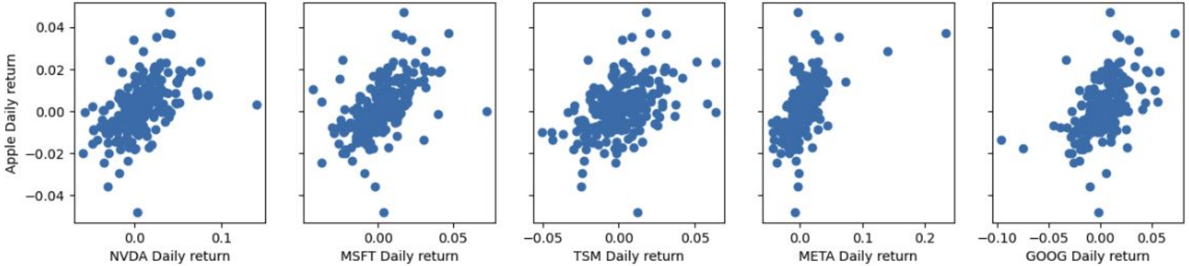
Regression on daily returns

$$DR = \frac{X_t - X_{t-1}}{X_{t-1}}$$

Years 2020-2022



Year 2023



Granger Causality determines VAR order

$$Y_t = \sum_{i=0}^{k} \beta_i Y_{t-i} + \epsilon$$

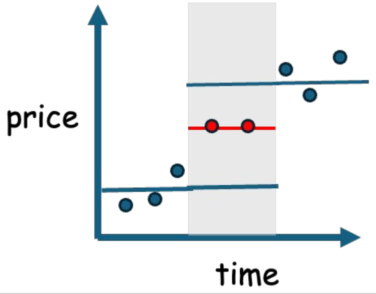
$$Y_t = \sum_{i=0}^{k} \beta_i Y_{t-i} + \sum_{j=0}^{l} \beta_i X_{t-j} + \eta$$

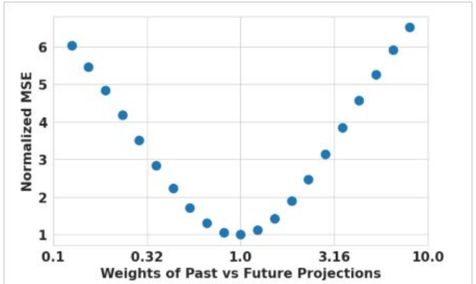
$$\overset{\text{GC}}{\Rightarrow} \eta < \epsilon$$

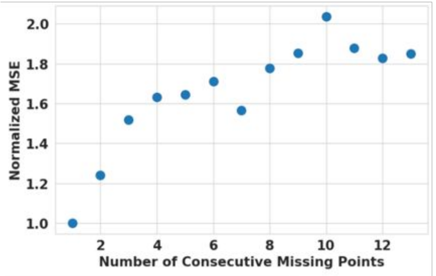
Y \ X	Apple	Google	NVIDIA
Apple	1.0	0.5876	0.0158
Google	0.0713	1.0	0.2592
NVIDIA	0.6800	0.8959	1.0

Results

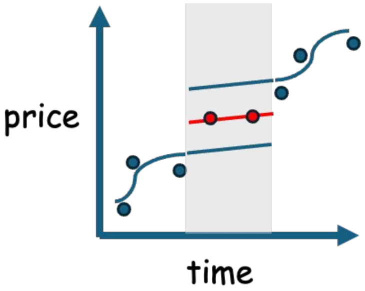
Rolling Average

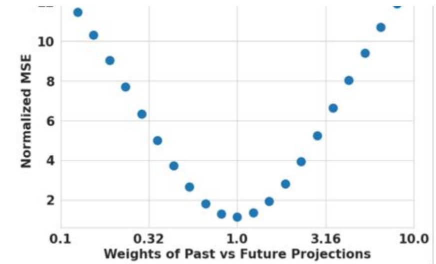


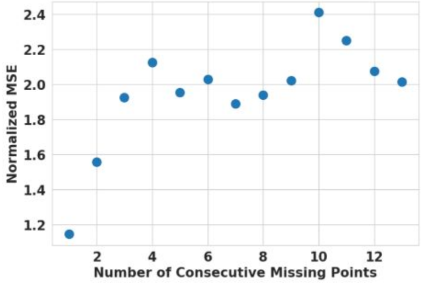




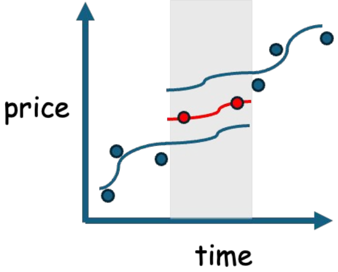
Double Exponential Smoothing

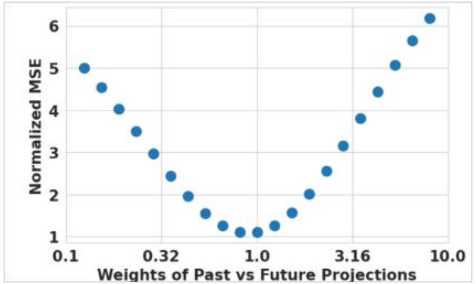


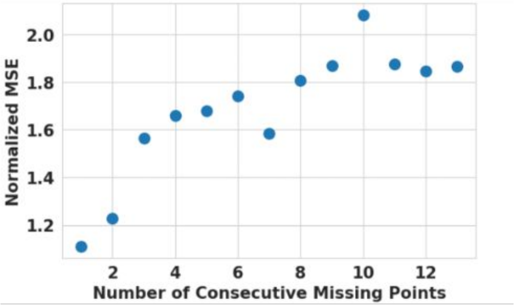




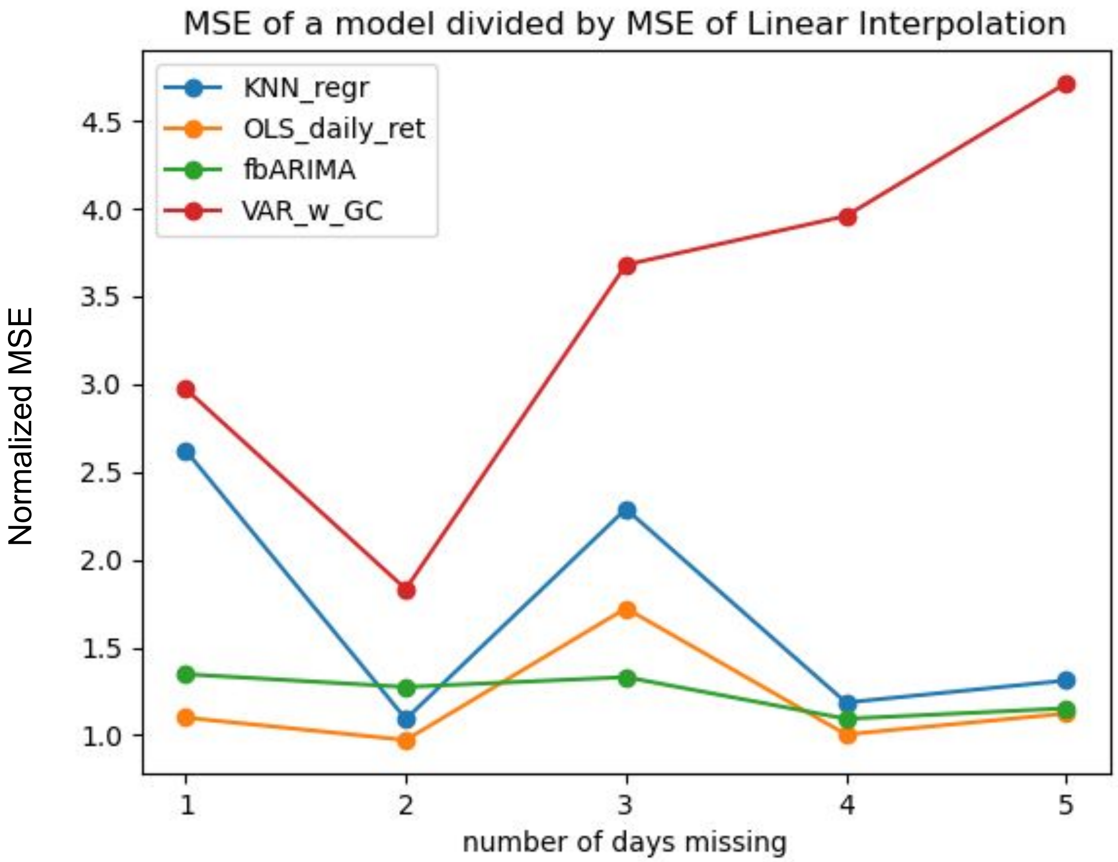
SARIMA







Regression, VAR and KNN, performance on year 2023 data.



Conclusion

- Linear interpolation is a robust choice for both small and large gaps.
- When there is sufficiently high correlation between prices movements of two companies, one may be used to impute missing data in the other.

Next steps

- Include other predictors that affect closing prices
- Systematically explore the circumstances under which the methods we evaluated outperform linear interpolation.
- Explore advanced techniques like State Space Models (Kalman Filter, Kalman Smoother) and Neural Network (MPL, Generative Adversarial Networks and Neural ODEs).

Acknowledgements

We would like to thank Roman Holowinsky, Steven Gubkin, Karthik Prabhu and Alec Clott