ITMO

Synthetic financial time series generation with regime clustering

Kirill Zakharov, Elizaveta Stavinova, Alexander Boukhanovsky
ITMO University, St Petersburg, RF

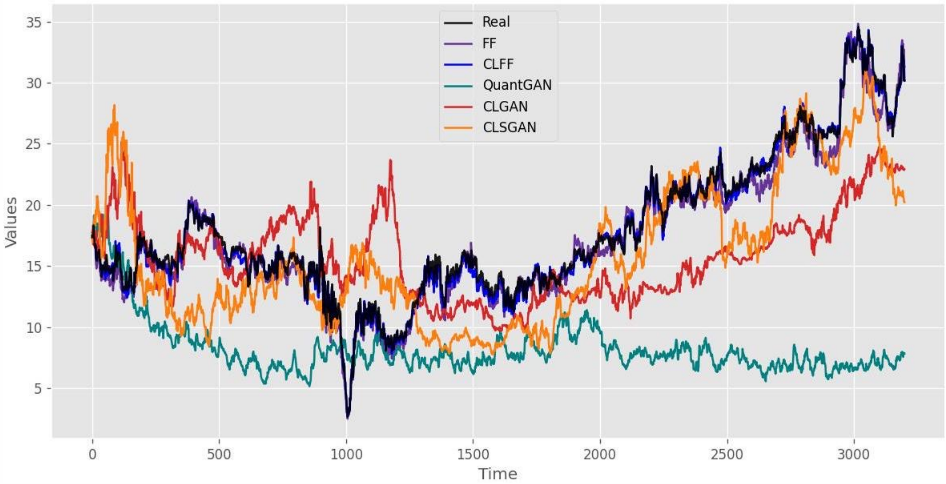
16th International Conference on Computer Science and Information Technology (ICCSIT 2023)
July 8, 2023
Paris, France

Why do we need synthetic data?

- Augmentation
- Scalability
- Privacy
- Filtration
- Quality improvement

Why do we need synthetic data?

- Augmentation
- Scalability
- Privacy
- Filtration
- Quality improvement



Specificity of financial time series

1. Reflect the dynamics of the redistribution of resources in the economy

expressed in terms of financial indicators

Specificity of financial time series

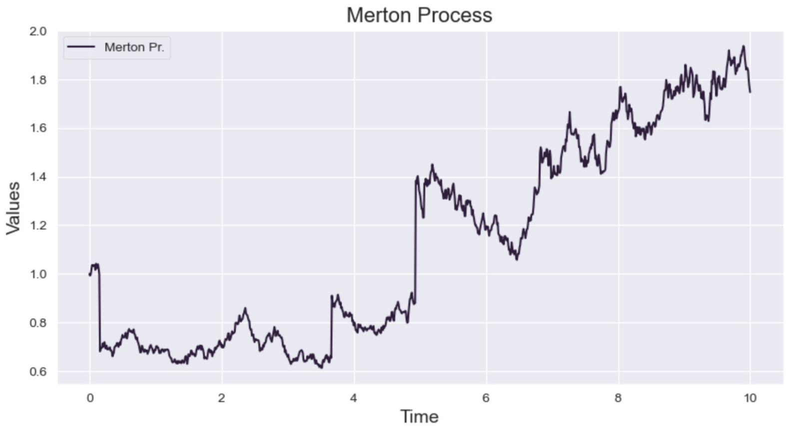
1. Reflect the dynamics of the redistribution of resources in the economy

2. Multi-scale and evolving nature

    - Weekly and seasonal patterns of behaviour along with the trend
    - The crisis impact

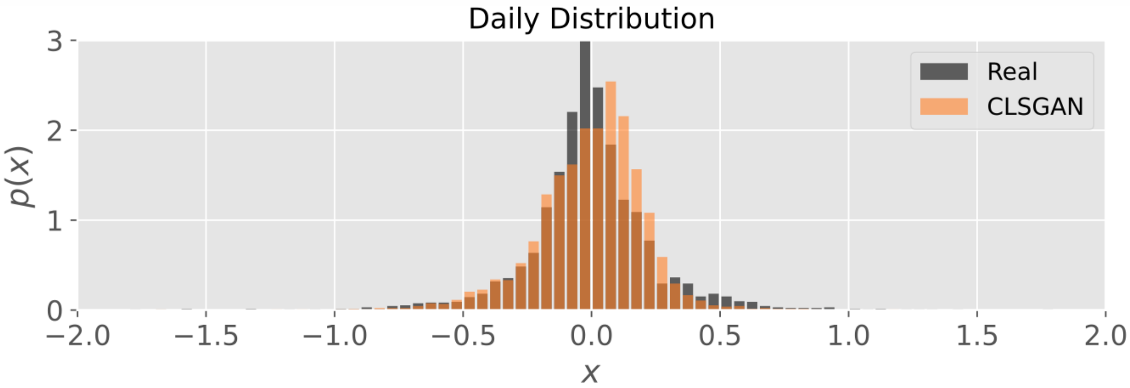
Specificity of financial time series

1. Reflect the dynamics of the redistribution of resources in the economy
2. Multi-scale and evolutionary character
3. Non-stationary, non-periodic
4. With erratic transitions between states



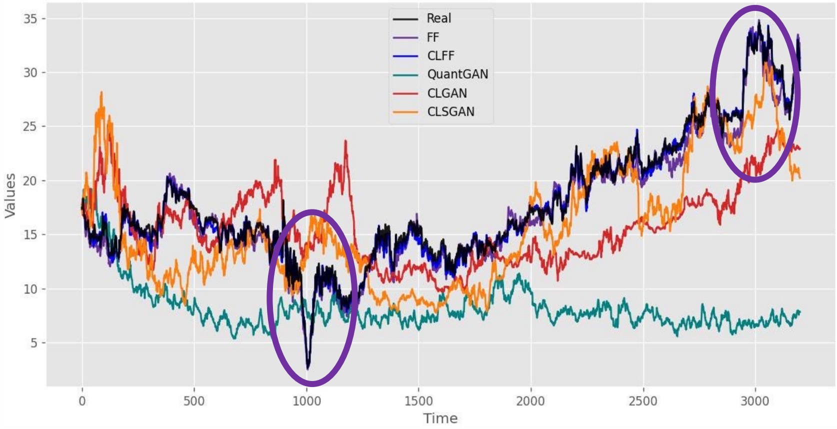
Specificity of financial time series

1. Reflect the dynamics of the redistribution of resources in the economy
2. Multi-scale and evolutionary character
3. Non-stationary, non-periodic
4. With erratic transitions between states
5. Heavy tails of log return distribution

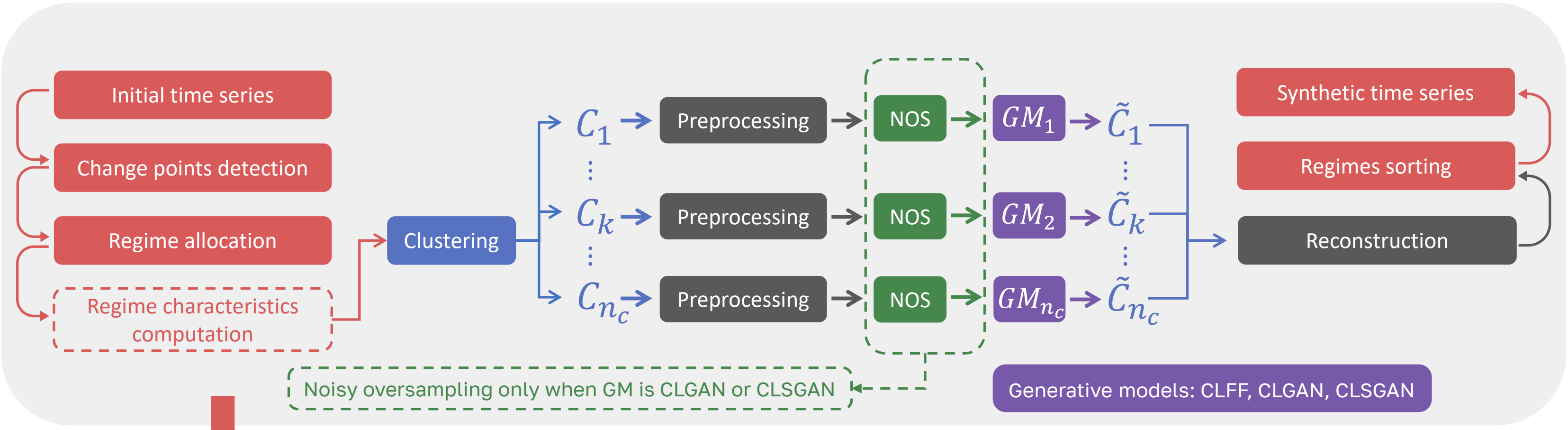


Specificity of financial time series

1. Reflect the dynamics of the redistribution of resources in the economy
2. Multi-scale and evolutionary character
3. Non-stationary, non-periodic
4. With erratic transitions between states
5. Heavy tails of log return distribution
6. Volatility clustering



Proposed method



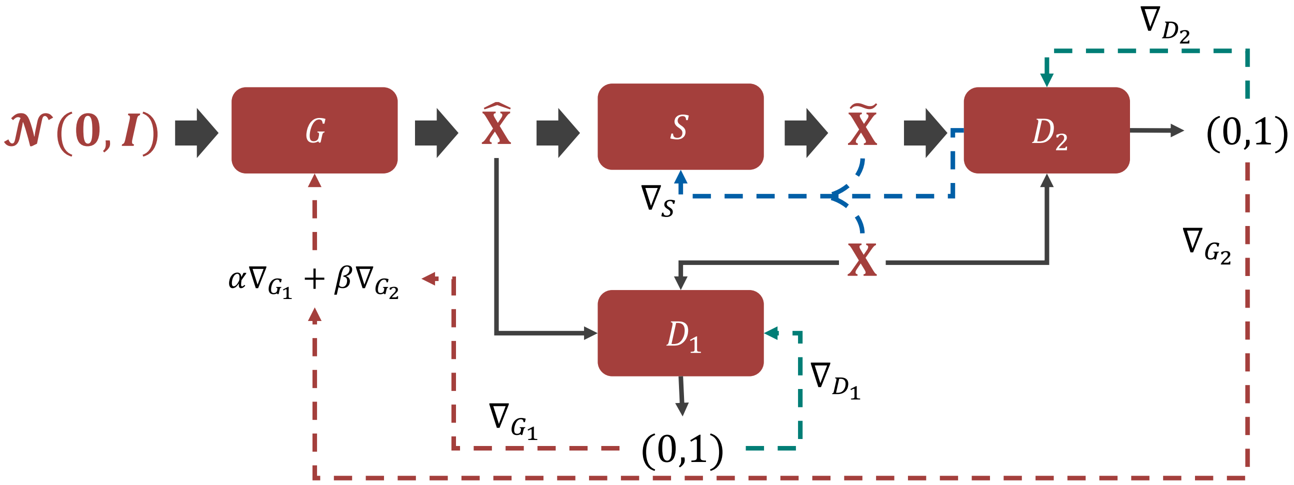
General pipeline

Characteristics:

- Mean
- Standard deviation
- Skewness
- Kurtosis
- Kolmogorov-Smirnov statistic
- Spectral density
- Min and max values

Generative models

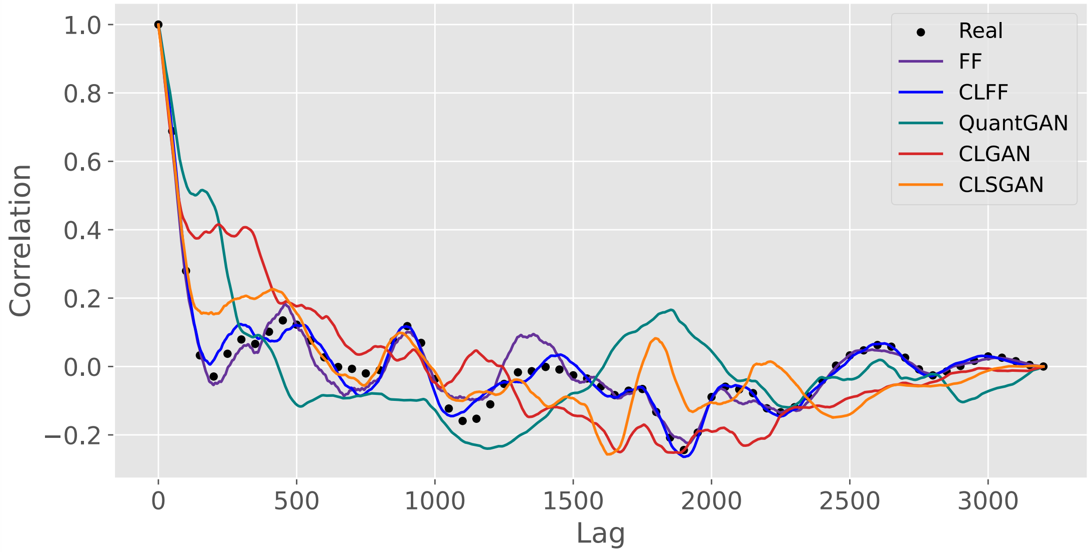
1. CLFF (modification of the Fourier Flows)
2. CLGAN (modification of the QuantGAN)
3. CLSGAN (new stable architecture)



CLSGAN training process

Experimental study: data and ACF

As data we use the open-source stock prices: GEN, ZEUS, FISI
These time series have approximately 3000 records

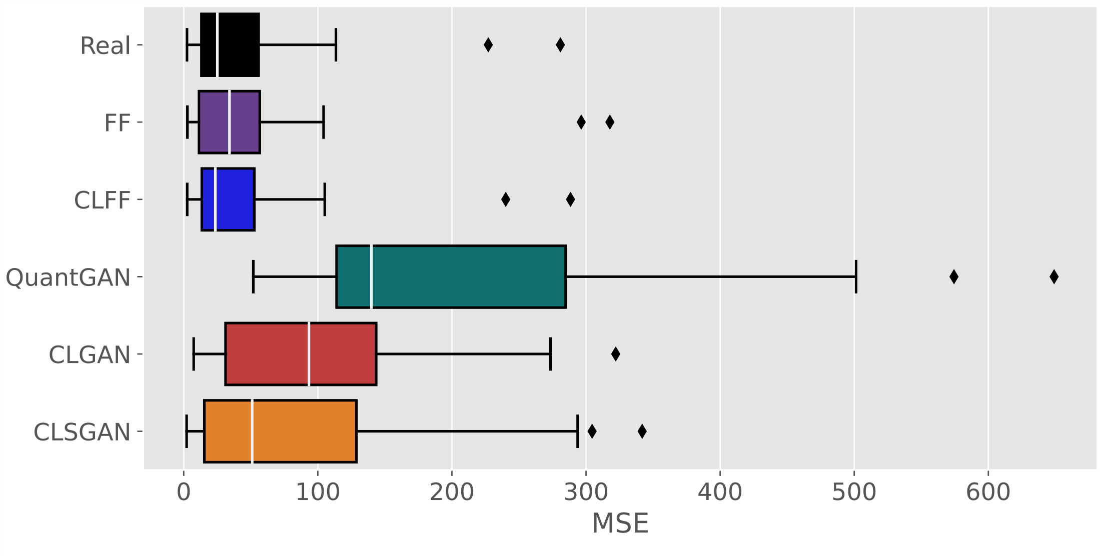


Autocorrelation function

- Our approach better approximates the autocorrelation patterns
- It can be used to examine historical price movements and predict future ones

Experimental study: forecasting model enhancement

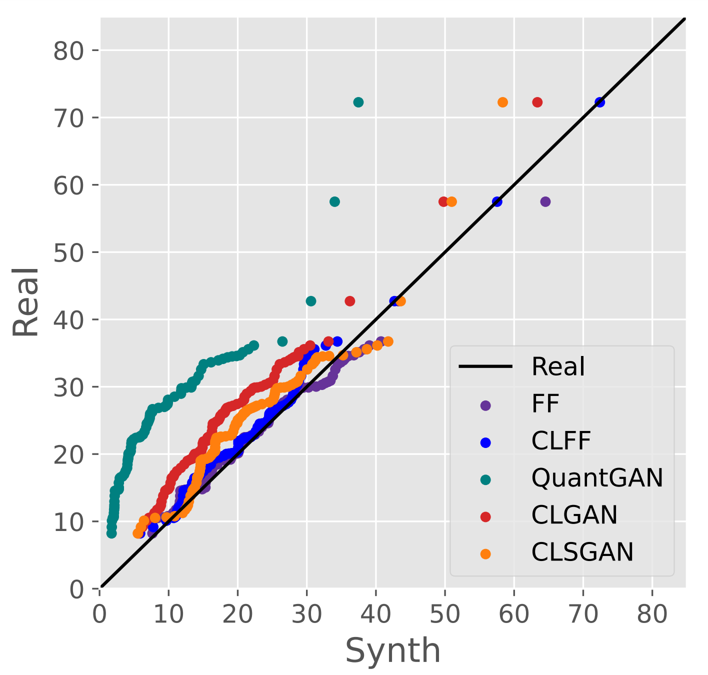
1. Train forecasting model on synthetic time series
2. Test it on the real-world time series using cross validation



Box plots of MSE errors

- In the case of CLFF we got less errors than on training on a real-world time series
- Therefore, we can use CLFF to change initial time series on synthetic one to make precise predictions

Experimental study: local extremum points consistency



Q-Q plots of Local extremum points

- Our approach better approximates the location of local extremum points in the initial time series
- Thus, you can use it in the tasks when important to get the precise dynamic of the initial time series

Quantitative analysis

Data	Methods	Skewness	Kurtosis	D_{JS}	S_{χ}	KS*
GEN	Real	**0.747**	**-0.135**	**0.000**	**2.004**	**0.000**
	FF	0.595 (±0.05)	-0.280 (±0.08)	0.077	2.350	0.159
	CLFF	**0.733** (±0.01)	**-0.152** (±0.01)	**0.017**	**2.001**	**0.029**
	QuantGAN	0.211 (±0.13)	**-0.393** (±0.18)	0.241	2.287	0.843
	CLGAN	**0.371** (±0.21)	-0.576 (±0.29)	0.162	2.256	**0.456**
	CLSGAN	0.336 (±0.31)	-0.552 (±0.41)	**0.141**	**1.995**	0.501
ZEUS	Real	**1.750**	**6.268**	**0.000**	**1.951**	**0.000**
	FF	0.996 (±0.15)	2.732 (±0.75)	0.048	2.434	0.347
	CLFF	**1.681** (±0.04)	**5.673** (±0.19)	**0.027**	**2.014**	**0.121**
	QuantGAN	0.548 (±0.21)	0.159 (±0.69)	0.171	2.084	0.426
	CLGAN	0.123 (±0.33)	-0.213 (±0.51)	0.192	2.107	0.282
	CLSGAN	**0.590** (±0.31)	**0.701** (±0.76)	**0.169**	**2.052**	**0.211**
FISI	Real	**0.735**	**0.198**	**0.000**	**2.072**	**0.000**
	FF	0.868 (±0.03)	0.601 (±0.09)	0.050	2.614	0.221
	CLFF	**0.739** (±0.01)	**0.238** (±0.02)	**0.020**	**2.102**	**0.039**
	QuantGAN	0.563 (±0.22)	**0.151** (±0.65)	0.180	2.157	0.467
	CLGAN	0.252 (±0.11)	-0.473 (±0.25)	0.185	1.954	0.354
	CLSGAN	**0.778** (±0.19)	-0.401 (±0.29)	**0.157**	**2.015**	**0.335**

Blue color – best results of NF based models
Orange color – best results of GAN based models

Conclusion

- We proposed a new method for synthetic financial time series generation, which can be used in situations where inferring an explicit probabilistic time series model is difficult
- Due to regimes clustering, our method can deal with multiscale nature of time series and generate a data containing a diversity of patterns presented in the initial data
- We proposed three generative models, that can be used inside the method depending on the desired quality of the synthetic data: CLFF results in an accurate data generation, while CLGAN and CLSGAN provide generation of a more diverse data
- The developed method can be applied to the tasks of historical data supplementation for training a ML model of a desired quality or historical data replacement in case of data sharing restrictions

THANK YOU FOR YOUR TIME!

ITsMOre than a UNIVERSITY

kazakharov@itmo.ru