Are Transformers Effective for Time Series Forecasting?

Ailing Zeng, Muxi Chen (2022)

Presenter: Jae-Hoon Kim and Yong-Won Seo

November 27, 2023

Contents

1. Time series forecasting

Time series forecasting

- For time series containing C variables, given historical data

$$\mathcal{X} = \left\{X_1^t, \ldots, X_C^t\right\}_{t=1}^{L}, \ \textit{where}$$

- L is the look-back window size.
- X_i^t is the value of the i_{th} variate at the t_{th} time step.

Time series forecasting

- The time series forecasting task is to predict the values

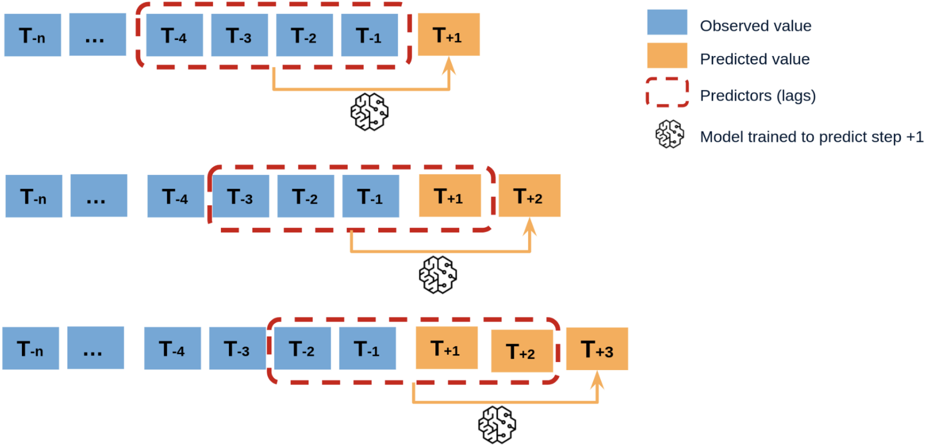
$$\hat{\mathcal{X}} = \left\{ \hat{X}_1^t, \ldots, \hat{X}_C^t \right\}_{t=L+1}^{L+T}, \ \textit{where}$$

- T is the number of future steps.
- Two types of time series forecasting
    - IMS: Iterated multi-step forecasting
    - DMS: Direct multi-step forecasting

IMS forecasting

- IMS forecasting learns a single step forecaster and iteratively applies it to obtain multi-step predictions.
- The general properties of IMS forecasting is that
  - have smaller variance.
  - have error accumulation effects.
- Thus, it is preferable
  - with a highly-accurate single-step forecaster.
  - when T is relatively small.

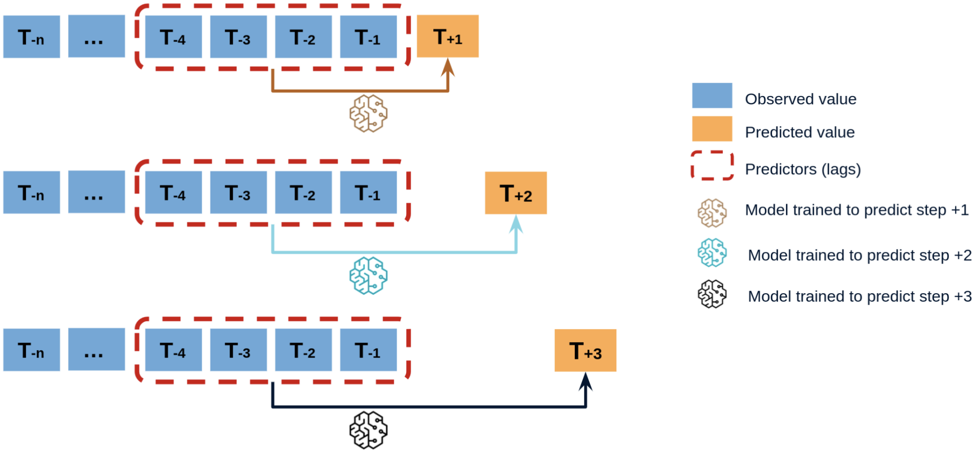
IMS forecasting: Visualization



DMS forecasting

- DMS forecasting directly optimizes the multi-step forecasting objective at once.
- It generates more accurate predictions when
  - obtaining an unbiased single-step forecasting model is hard.
  - T is relatively large.
- In long-term time series forecasting, DMS forecasting generally shows better performance
- Most of the existing transformer models improve their performances due to this strategy.

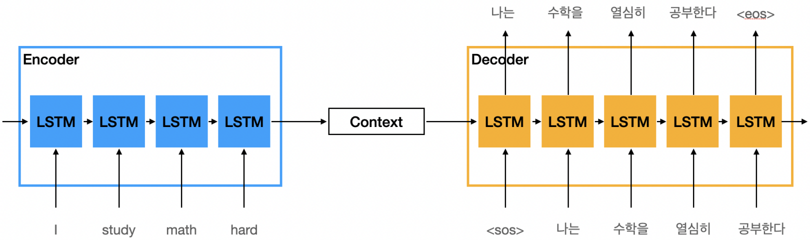
DMS forecasting: Visualization



2. The Transformer models

The Seq2Seq models

- The Transformer models are more advanced and efficient alternative to the traditional Seq2Seq models.
- The limitation of the traditional Seq2Seq models
  - fixed context vector dimension acts as a bottleneck
  - gradient vanishing problem in the context of long sequences



The Transformer models

- The Transformer models solved these problems using
  - Positional encoding techniques
  - Multi-head self attention mechanisms
- Become the most successful sequence modeling architecture demonstrating unparalleled performance
- Surge of Transformer-based solutions for time series analysis with these notable models: FED-former, Autoformer, Informer, $\cdots$

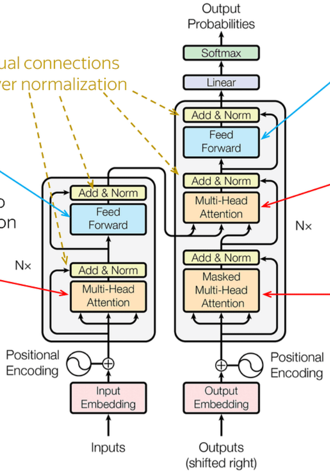
The Transformer models: Visualization

Residual connections
and layer normalization

Feed-forward network:
after taking information from
other tokens, take a moment to
think and process this information

Encoder self-attention:
tokens look at each other

queries, keys, values
are computed from
encoder states



Feed-forward network:
after taking information from
other tokens, take a moment to
think and process this information

Decoder-encoder attention:
target token looks at the source

queries – from decoder states; keys
and values from encoder states

Decoder self-attention (masked):
tokens look at the previous tokens

queries, keys, values are computed
from decoder states

Self-attention mechanism

- Main working power of transformer models
- Extracting semantic correlations among elements in a long sequence
- Inevitable to have temporal information loss despite positional encoding techniques
- Especially serious for Time series data since
  - Lack of semantics in the numerical data itself
  - The order plays the most crucial role

3. LTSF-Linear model

LTSF-Linear model

- The writer proposes the model

  $\hat{X}_i = WX_i$, where $W \in \mathbb{R}^{T \times L}$ is a linear layer among the temporal axis.
- Two variants with two pre-processing methods: DLinear and NLinear models.

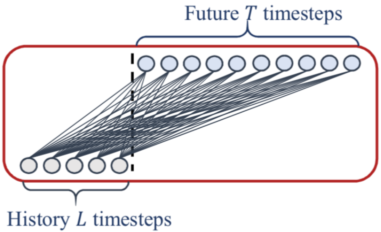
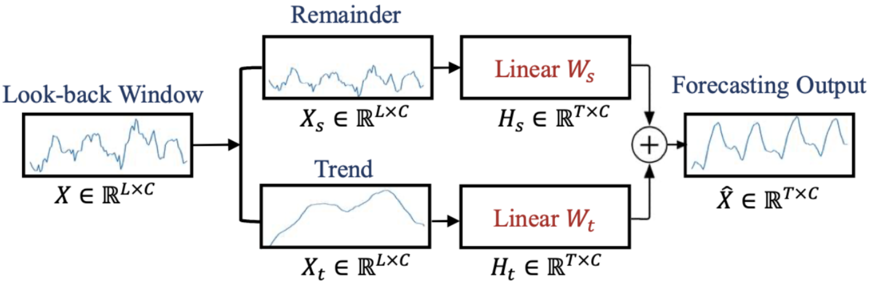


Figure 2. Illustration of the basic linear model.

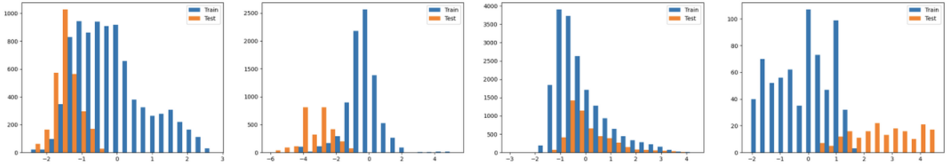
DLinear model

- The combination of the decomposition scheme used in Autoformer and FEDformer.
- Using a moving average kernel, decompose raw data into
  - a trend component
  - a remainder (seasonal) component
- DLinear enhances the performance when there is a clear trend in the data



NLinear model

- Boost the performance when there is a distribution shift in the dataset by
    1. Normalizing input data by subtracting them by the last value
    2. The input goes through a linear layer
    3. The subtracted part is added back



(a) ETTh1 channel6 (b) ETTh2 channel3 (c) Electricity channel3 (d) ILI channel6

Figure 5. Distribution of ETTh1, ETTh2, Electricity, and ILI dataset. A clear distribution shift between training and testing data can be observed in ETTh1, ETTh2, and ILI.

4. Experiments

Data description

- The writer conducts experiments on nine widely-used real-world datasets, including ETT (Electricity Transformer Temperature), Traffic, Electricity, Weather, ILI, and Exchange rate.
- Detailed information such as the number of variables, timesteps and granularity is

Datasets	ETTh1&ETTh2	ETTm1 &ETTm2	Traffic	Electricity	Exchange-Rate	Weather	ILI
Variates	7	7	862	321	8	21	7
Timesteps	17,420	69,680	17,544	26,304	7,588	52,696	966
Granularity	1hour	5min	1hour	1hour	1day	10min	1week

Table 1. The statistics of the nine popular datasets for the LTSF problem.

Experiment settings

- The writer compares the models
  - Linear methods: LTSF-Linear, NLinear, DLinear.
  - Transformers: FED-former, Autoformer, Informer, Pyraformer, and LogTrans.
  - Closest Repeat (Repeat): repeats the last value in the look-back window.

- The writer uses Mean Squared Error(MSE) and Mean Absolute Error(MAE) as the core metrics to compare performance.
- The writer uses various forecasting horizon $T \in \{96, 192, 336, 720\}$

Performances comparison

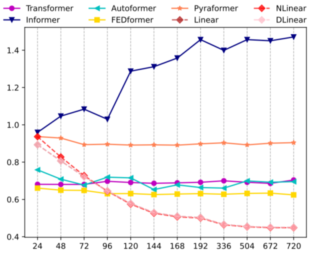
Methods		IMP.	Linear*		NLinear*		DLinear*		FEDformer		Autoformer		Informer		Pyraformer*		LogTrans		*Repeat**	
Metric		MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	27.40%	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**	0.193	0.308	0.201	0.317	0.274	0.368	0.386	0.449	0.258	0.357	1.588	0.946
	192	23.88%	**0.153**	0.250	0.154	**0.248**	**0.153**	0.249	0.201	0.315	0.222	0.334	0.296	0.386	0.386	0.443	0.266	0.368	1.595	0.950
	336	21.02%	**0.169**	0.268	0.171	**0.265**	**0.169**	0.267	0.214	0.329	0.231	0.338	0.300	0.394	0.378	0.443	0.280	0.380	1.617	0.961
	720	17.47%	**0.203**	0.301	0.210	**0.297**	**0.203**	0.301	0.246	0.355	0.254	0.361	0.373	0.439	0.376	0.445	0.283	0.376	1.647	0.975
Exchange	96	45.27%	0.082	0.207	0.089	0.208	**0.081**	0.203	0.148	0.278	0.197	0.323	0.847	0.752	0.376	1.105	0.968	0.812	**0.081**	**0.196**
	192	42.06%	0.167	0.304	0.180	0.300	**0.157**	0.293	0.271	0.380	0.300	0.369	1.204	0.895	1.748	1.151	1.040	0.851	0.167	**0.289**
	336	33.69%	0.328	0.432	0.331	0.415	**0.305**	0.414	0.460	0.500	0.509	0.524	1.672	1.036	1.874	1.172	1.659	1.081	**0.305**	**0.396**
	720	46.19%	0.964	0.750	1.033	0.780	**0.643**	**0.601**	1.195	0.841	1.447	0.941	2.478	1.310	1.943	1.206	1.941	1.127	0.823	0.681
Traffic	96	30.15%	**0.410**	0.282	**0.410**	**0.279**	**0.410**	0.282	0.587	0.366	0.613	0.388	0.719	0.391	2.085	0.468	0.684	0.384	2.723	1.079
	192	29.96%	**0.423**	0.287	**0.423**	**0.284**	**0.423**	0.287	0.604	0.373	0.616	0.382	0.696	0.379	0.867	0.467	0.685	0.390	2.756	1.087
	336	29.95%	0.436	0.295	**0.435**	**0.290**	0.436	0.296	0.621	0.383	0.622	0.337	0.777	0.420	0.869	0.469	0.734	0.408	2.791	1.095
	720	25.87%	0.466	0.315	**0.464**	**0.307**	0.466	0.315	0.626	0.382	0.660	0.408	0.864	0.472	0.881	0.473	0.717	0.396	2.811	1.097
Weather	96	18.89%	**0.176**	0.236	0.182	**0.232**	**0.176**	0.237	0.217	0.296	0.266	0.336	0.300	0.384	0.896	0.556	0.458	0.490	0.259	0.254
	192	21.01%	**0.218**	0.276	0.225	**0.269**	0.220	0.282	0.276	0.336	0.307	0.367	0.598	0.544	0.622	0.624	0.658	0.589	0.309	0.292
	336	22.71%	**0.262**	0.312	0.271	**0.301**	0.265	0.319	0.339	0.380	0.359	0.395	0.578	0.523	0.739	0.753	0.797	0.652	0.377	0.338
	720	19.85%	0.326	0.365	0.338	**0.348**	**0.323**	0.362	0.403	0.428	0.419	0.428	1.059	0.741	1.004	0.934	0.869	0.675	0.465	0.394
ILI	24	47.86%	1.947	0.985	**1.683**	**0.858**	2.215	1.081	3.228	1.260	3.483	1.287	5.764	1.677	1.420	2.012	4.480	1.444	6.587	1.701
	36	36.43%	2.182	1.036	**1.703**	**0.859**	1.963	0.963	2.679	1.080	3.103	1.148	4.755	1.467	7.394	2.031	4.799	1.467	7.130	1.884
	48	34.43%	2.256	1.060	**1.719**	**0.884**	2.130	1.024	2.622	1.078	2.669	1.085	4.763	1.469	7.551	2.057	4.800	1.468	6.575	1.798
	60	34.33%	2.390	1.104	**1.819**	**0.917**	2.368	1.096	2.857	1.157	2.770	1.125	5.264	1.564	7.662	2.100	5.278	1.560	5.893	1.677

Performances comparison(Cond.)

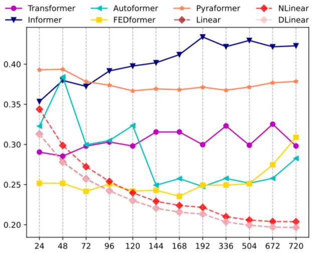
ETTh1	96	0.80%	0.375	0.397	**0.374**	**0.394**	0.375	0.399	0.376	0.419	0.449	0.459	0.865	0.713	0.664	0.612	0.878	0.740	1.295	0.713
	192	3.57%	0.418	0.429	0.408	**0.415**	**0.405**	0.416	0.420	0.448	0.500	0.482	1.008	0.792	0.790	0.681	1.037	0.824	1.325	0.733
	336	6.54%	0.479	0.476	**0.429**	**0.427**	0.439	0.443	0.459	0.465	0.521	0.496	1.107	0.809	0.891	0.738	1.238	0.932	1.323	0.744
	720	13.04%	0.624	0.592	**0.440**	**0.453**	0.472	0.490	0.506	0.507	0.514	0.512	1.181	0.865	0.963	0.782	1.135	0.852	1.339	0.756
ETTh2	96	19.94%	0.288	0.352	**0.277**	**0.338**	0.289	0.353	0.346	0.388	0.358	0.397	3.755	1.525	0.645	0.597	2.116	1.197	0.432	0.422
	192	19.81%	0.377	0.413	**0.344**	**0.381**	0.383	0.418	0.429	0.439	0.456	0.452	5.602	1.931	0.788	0.683	4.315	1.635	0.534	0.473
	336	25.93%	0.452	0.461	**0.357**	**0.400**	0.448	0.465	0.496	0.487	0.482	0.486	4.721	1.835	0.907	0.747	1.124	1.604	0.591	0.508
	720	14.25%	0.698	0.595	**0.394**	**0.436**	0.605	0.551	0.463	0.474	0.515	0.511	3.647	1.625	0.963	0.783	3.188	1.540	0.588	0.517
ETTm1	96	21.10%	0.308	0.352	0.306	0.348	**0.299**	**0.343**	0.379	0.419	0.505	0.475	0.672	0.571	0.543	0.510	0.600	0.546	1.214	0.665
	192	21.36%	0.340	0.369	0.349	0.375	**0.335**	**0.365**	0.426	0.441	0.553	0.496	0.795	0.669	0.557	0.537	0.837	0.700	1.261	0.690
	336	17.07%	0.376	0.393	0.375	0.388	**0.369**	**0.386**	0.445	0.459	0.621	0.537	1.212	0.871	0.754	0.655	1.124	0.832	1.283	0.707
	720	21.73%	0.440	0.435	0.433	0.422	**0.425**	**0.421**	0.543	0.490	0.671	0.561	1.166	0.823	0.908	0.724	1.153	0.820	1.319	0.729
ETTm2	96	17.73%	0.168	0.262	**0.167**	**0.255**	**0.167**	0.260	0.203	0.287	0.255	0.339	0.365	0.453	0.435	0.507	0.768	0.642	0.266	0.328
	192	17.84%	0.232	0.308	**0.221**	**0.293**	0.224	0.303	0.269	0.328	0.281	0.340	0.533	0.563	0.730	0.673	0.989	0.757	0.340	0.371
	336	15.69%	0.320	0.373	**0.274**	**0.327**	0.281	0.342	0.325	0.366	0.339	0.372	1.363	0.887	1.201	0.845	1.334	0.872	0.412	0.410
	720	12.58%	0.413	0.435	**0.368**	**0.384**	0.397	0.421	0.421	0.415	0.433	0.432	3.379	1.338	3.625	1.451	3.048	1.328	0.521	0.465

Performances regarding look-back sizes

- A powerful TSF model with a strong temporal relation extraction capability should be able to achieve better results with larger look-back window sizes.
- Unlike Linear models, existing Transformer-based models' performance deteriorates or stays stable when the look-back window size increases.



(a) **720** steps-**Traffic**



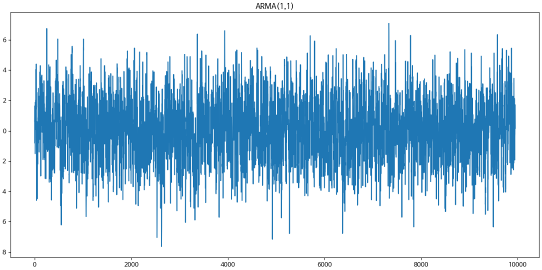
(b) **720** steps-**Electricity**

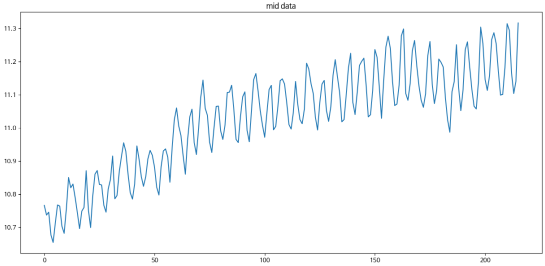
5. Forecasting with LTSF-Linear models

Data analysis

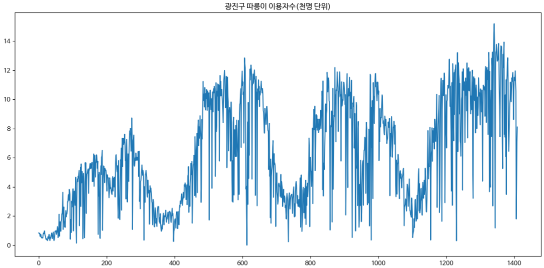
- Time series data sampled from an ARMA(1,1) model.
- Data provided in the midterm examination.
- Daily usage data of Seoul's public bike, TTareungyi, in Gwangjin-gu.
- Data on Won-Dollar exchange rates since the 1990s.

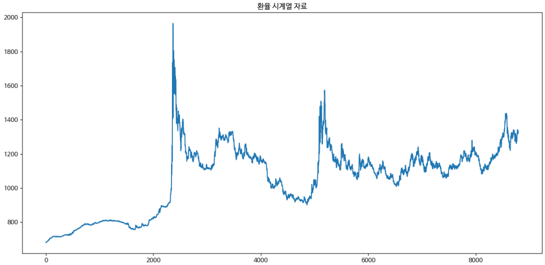
Data analysis



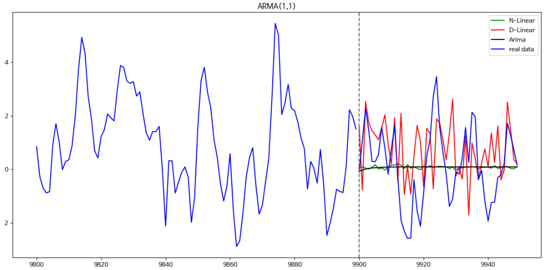


Data analysis



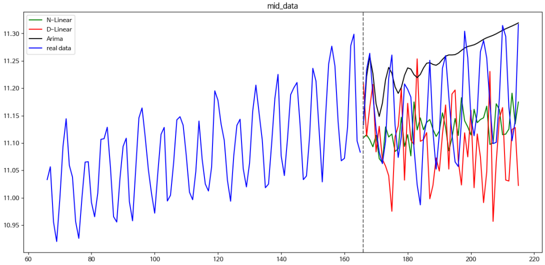


ARMA(1,1)



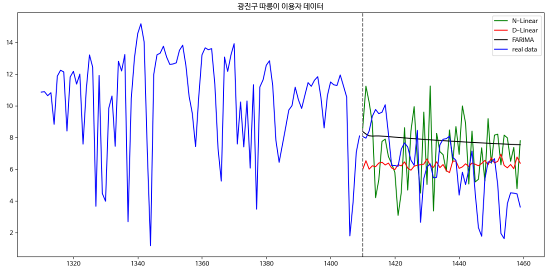
	N 100	N 200	N 300	D 100	D 200	D 300	ARIMA
MSE	4.357	4.352	4.635	2.764	2.754	2.727	2.638
MAE	1.592	1.665	1.729	1.291	1.268	1.298	1.251

Midterm



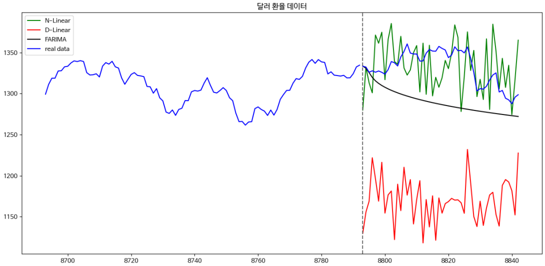
	N 50	N 100	D 50	D 100	ARIMA
MSE	0.0181	0.0178	0.0111	0.0088	0.0133
MAE	0.1067	0.1043	0.0838	0.0726	0.0914

Bicycle

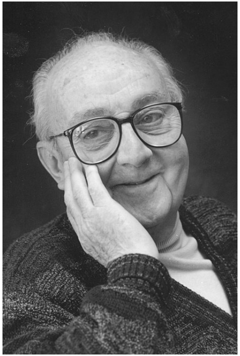


	N 100	N 200	N 300	D 100	D 200	D 300	FARIMA
MSE	10.8722	8.5861	10.7723	5.8166	5.4810	4.9605	6.4405
MAE	2.6724	2.2851	2.5849	1.9026	1.7908	1.6977	1.9932

Exchange90



	L 100	L 200	L 300	D 100	D 200	D 300	FARIMA
MSE	2382.25	2195.04	1544.35	60014.19	32784.19	19531.17	2181.0359
MAE	41.75	38.64	30.34	242.44	177.83	135.49	41.2548



”All models are wrong, but some are useful”.

-George Box-