西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables

ICLR 2024

主讲人：张玉

2024.04.03

目录
CONTENTS

01 Introduction

Introduction-Variable

内生变量(endogenous variables)：只关注感兴趣的目标

外生变量(exogenous variables)：为内生变量提供有价值的外部信息

通过外生变量引入辅助信息来促进内生变量的预测

Example(from datasets):

(1)ECL：客户的每小时用电量数据。

将最后一个客户端的用电量作为内生变量，其他客户端作为外生变量。

(2)天气：气象站每10分钟收集的21个气象因子。

使用Wet Bulb因子作为内生变量，其他指标作为外生变量。

(3)ETT：变压器油温数据。

内生变量为油温，外生变量为6个电力负荷特征。

(4)交通：高速公路传感器测量的每小时道路占用率。

将最后一个传感器的测量作为内生变量，将其他传感器作为外生变量。

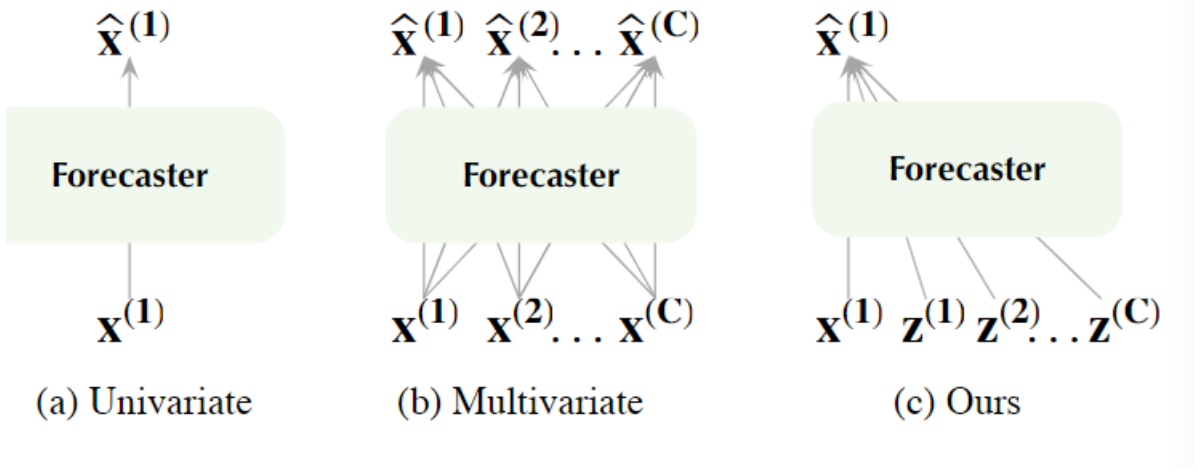
Introduction-Question

内生变量和外生变量之间：需要调和差异和依赖性

外部因素对内生序列的影响：可能是连续和时滞的

主要参考：

- patchTST :只能捕获时间依赖性，不能捕获多变量之间的相关性（通道独立）
- itransformer:无法捕获不同子序列之间的时间变化（一个series对于一个token)



(a) Univariate (b) Multivariate (c) Ours

Introduction-Advantage

TimeXer:

- 在包含外生变量的基础上，不修改transformer的架构(模型简单)
- self-attention:对Patch-level的内生时序token提取时间依赖性（相关性）
- cross-attention:对变量token提取多变量相关性（外生变量对内生变量的影响）

Introduction-Advantage

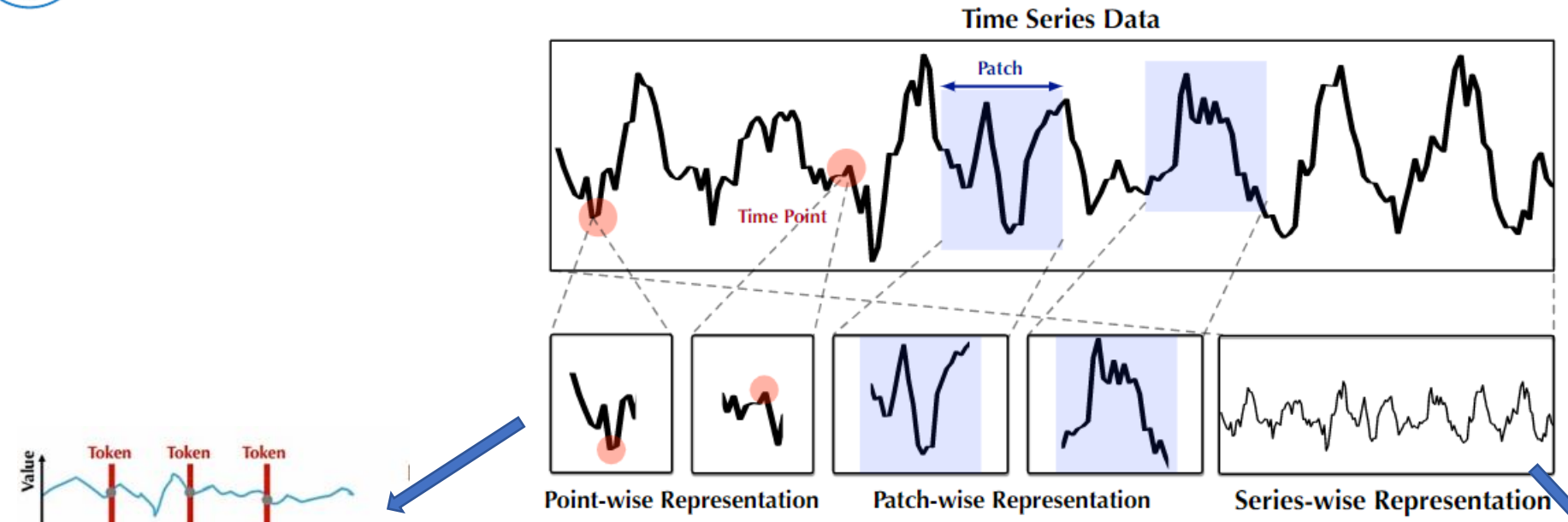
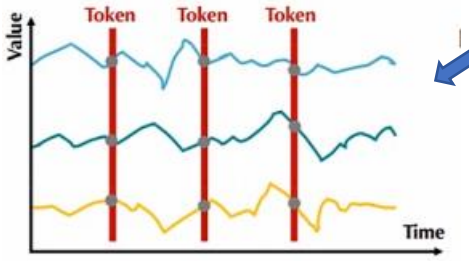
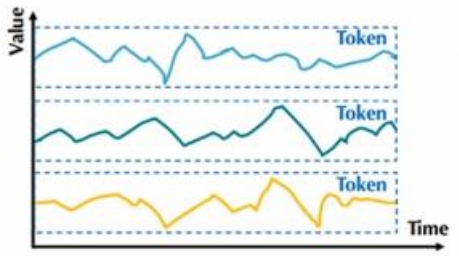


Figure 2. Illustration of different levels of representation for time series data, ranging from point-wise, patch-wise to series-wise.



Original Transformer

PatchTST



iTransformer

02 TimeXer Structure

Problem Definition

只预测内生时间序列，外生变量是附加因素

内生序列：$\mathbf{x}_{1:L} = \{x_1, x_2, ..., x_L\} \in \mathbb{R}^{L\times 1}$

外生序列：$\mathbf{z}_{1:L'} = \{\mathbf{z}_{1:L'}^{(1)}, \mathbf{z}_{1:L'}^{(2)}, ..., \mathbf{z}_{1:L'}^{(C)}\} \in \mathbb{R}^{L\times C}$

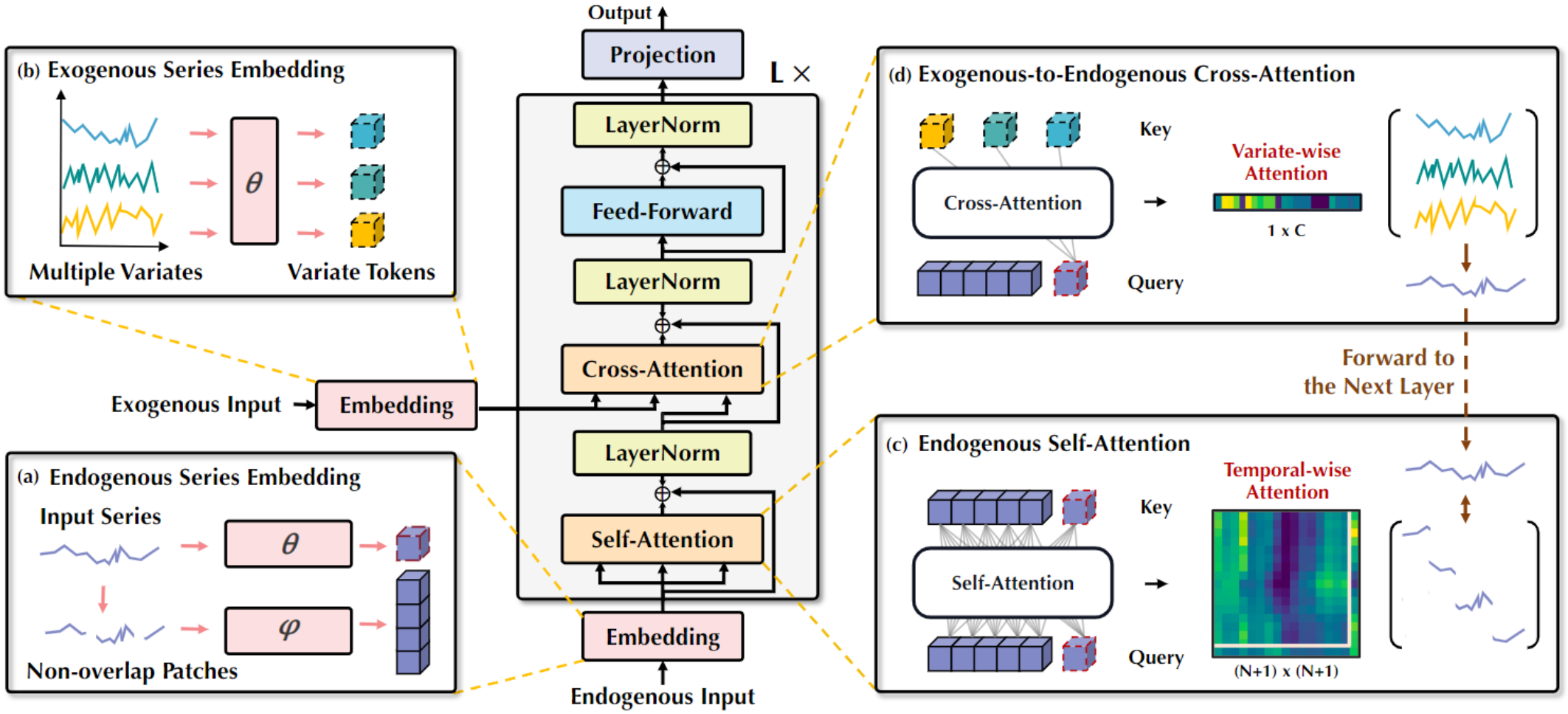
其中，内生序列和外生序列的长度可以不一致，即L和L`可以不相等

预测：

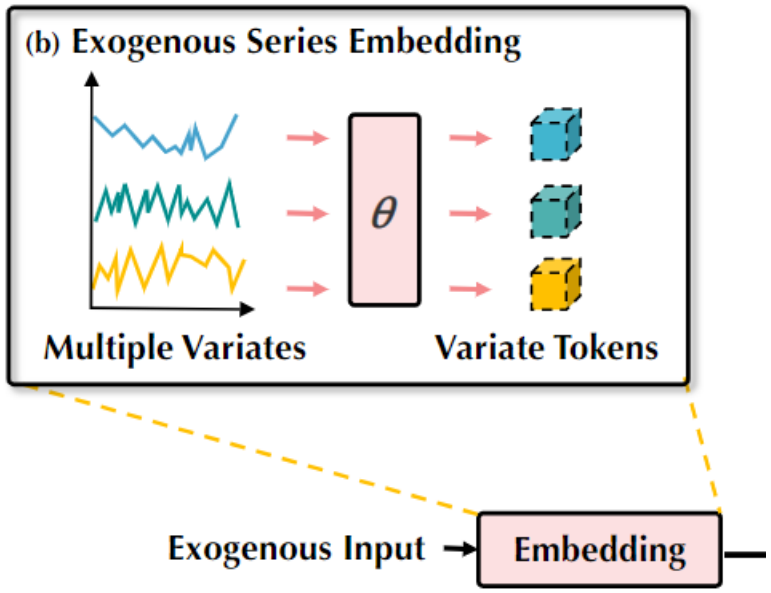
$$\widehat{\mathbf{x}} = f\left(\mathbf{x}_{1:L}, \mathbf{z}_{1:L'}\right).$$

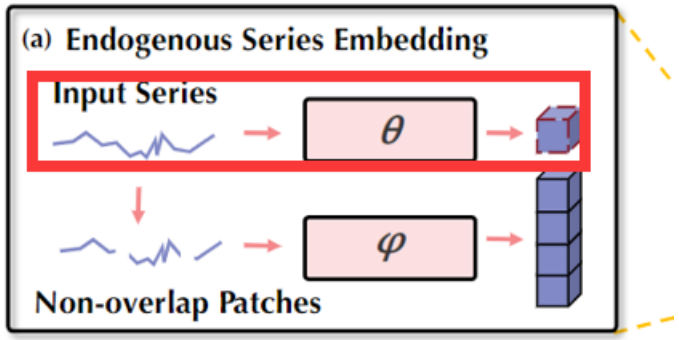
预测长度为S，即 $\widehat{\mathbf{x}} = \{x_{L+1}, x_{L+2}, ..., x_{L+S}\}$

Structure Overview



Variate Embedding





基于Itransformer的嵌入方式（线性层）

series->token 全局标记

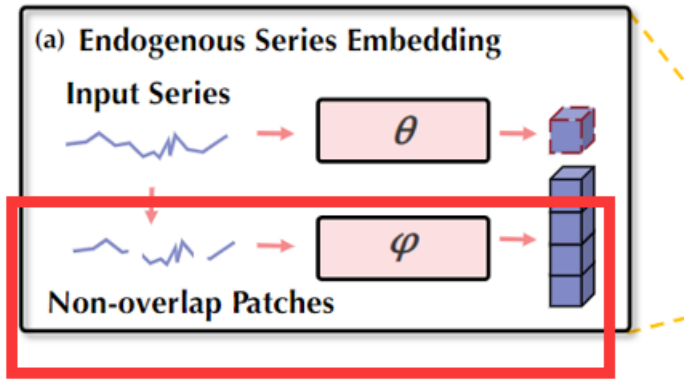
$$\mathbf{V}_{en} = \mathrm{EnVariateEmbed}\left(\mathbf{x}\right),$$
$$\mathbf{V}_{ex,i} = \mathrm{ExVariateEmbed}\left(\mathbf{z}^{(i)}\right)\ i \in \{1, \cdots, C\}.$$

EnVariateEmbed: $\mathbb{R}^L \rightarrow \mathbb{R}^D$

ExVariateEmbed: $\mathbb{R}^{L'} \rightarrow \mathbb{R}^D$

嵌入后的序列包含原始序列的全局信息

Patch Embedding



$$\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\} = \text{Patchify}(\mathbf{x}),$$
$$\mathbf{P}_{en} = \text{PatchEmbed}(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N).$$

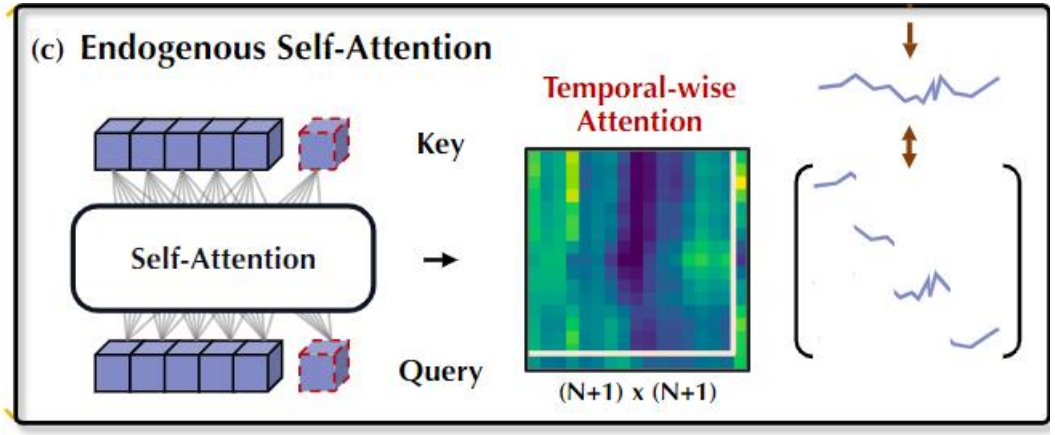
先将序列分为patch然后再嵌入为token

注意：此时采取的是不重叠的Patch块

对于内生变量而言，patch长度为P，N表示patch块的数量 PathEmbed函数将每个长度为P的patch转换为D维的嵌入

因此，得到的Pen的维度为NXD，N块D维度。

Patch-wise Self-Attention



Self-Attention+残差+LN归一化层:

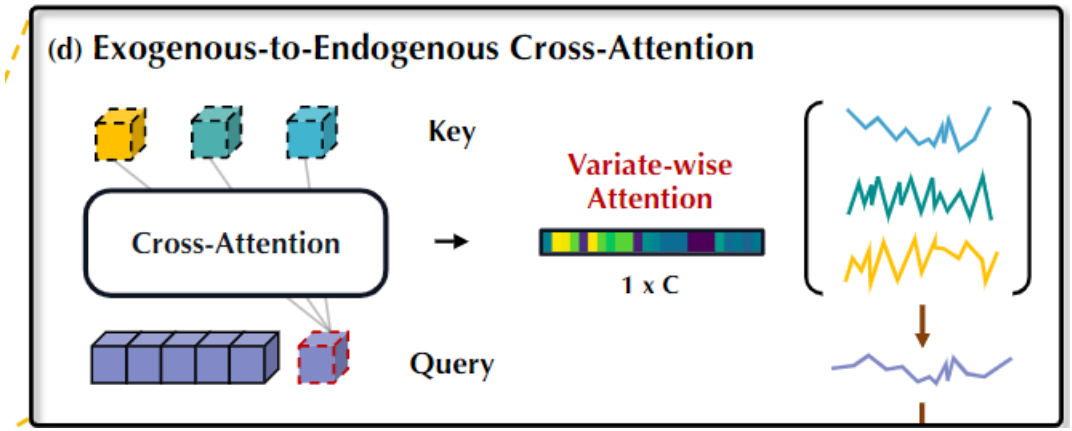
$$\widehat{\mathbf{P}}_{en}^{l}, \widehat{\mathbf{V}}_{en}^{l} = \mathrm{LN}\left(\left[\mathbf{P}_{en}^{l}, \mathbf{V}_{en}^{l}\right] + \text{Self-Attn}\left(\left[\mathbf{P}_{en}^{l}, \mathbf{V}_{en}^{l}\right]\right) \right)$$

[ ]表示将Pen和Ven拼接，此时维度变为N+1，N个patch,一个Series。

考虑到模型不需要预测外生变量，故只对内生变量做多头自注意来捕获内生变量之间的时间依赖性。

使用的内生变量为patch-wise的Pen和Series-wise的Ven,其中 Ven为整个内生序列提供了全局信息。

Variate-wise Cross-Attention



cross-attention+残差+LN归一化层

$$\mathbf{V}_{en}^{l+1} = \mathrm{LN}\left(\widehat{\mathbf{V}}_{en}^{l} + \text{Cross-Attn}\left(\widehat{\mathbf{V}}_{en}^{l}, \mathbf{V}_{ex}, \mathbf{V}_{ex}\right)\right)$$

使用交叉注意力来捕获外生序列与内生序列之间的相关性，其中，内生变量（series-wise的结果)作为查询（Q），外生变量（token）作为键和值（K,V)。

交叉注意力层的输出作为下一个Block的自注意层的输入。

03 Experiment

Experiment--Datasets

长时预测--用电量、天气、变压器油温、交通；一个内生，多个外生
短时预测--5个主要市场的电价短期预测数据集：电价作为内生变量，两个明显影响电价的外生变量

	Dataset	#Num	Ex. Descriptions	En. Descriptions	Sampling Frequency	Dataset Size
长时预测	Electricity	320	Electricity Consumption	Electricity Consumption	1 Hour	(18317, 2633, 5261)
	Weather	20	Climate Feature	CO2-Concentration	10 Minutes	(36792, 5271, 10540)
	ETTh	6	Power Load Feature	Oil Temperature	1 Hour	(8545, 2881, 2881)
	ETTm	6	Power Load Feature	Oil Temperature	15 Minutes	(34465, 11521, 11521)
	Traffic	861	Road Occupancy Rates	Road Occupancy Rates	1 Hour	(12185, 1757, 3509)
短时预测	NP	2	Grid Load, Wind Power	Nord Pool Electricity Price	1 Hour	(36500, 5219, 10460)
	PJM	2	System Load, SyZonal COMED load	Pennsylvania-New Jersey-Maryland Electricity Price	1 Hour	(36500, 5219, 10460)
	BE	2	Generation, System Load	Belgium's Electricity Price	1 Hour	(36500, 5219, 10460)
	FR	2	Generation, System Load	France's Electricity Price	1 Hour	(36500, 5219, 10460)
	DE	2	Wind power, Ampirion zonal load	German's Electricity Price	1 Hour	(36500, 5219, 10460)

Experiment--短期预测

输入长度-168，patch长度-24 预测长度-24

Model	TimeXer (Ours)		iTrans. (2023)		RLinear (2023)		PatchTST (2022)		Cross. (2022)		TiDE (2023)		TimesNet (2023a)		DLinear (2023)		SCINet (2022a)		Stationary (2022b)		Auto. (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	**0.238**	**0.268**	0.265	0.300	0.335	0.340	0.267	0.284	0.245	0.289	0.335	0.340	0.250	0.289	0.309	0.321	0.373	0.368	0.294	0.308	0.402	0.398
PJM	**0.088**	**0.188**	0.097	0.197	0.124	0.229	0.106	0.209	0.149	0.198	0.124	0.228	0.097	0.195	0.108	0.215	0.143	0.259	0.122	0.228	0.168	0.267
BE	**0.374**	**0.241**	0.394	0.270	0.520	0.337	0.403	0.264	0.436	0.294	0.523	0.336	0.419	0.288	0.463	0.313	0.731	0.412	0.433	0.289	0.500	0.333
FR	**0.381**	**0.211**	0.439	0.233	0.507	0.290	0.411	0.220	0.440	0.216	0.510	0.290	0.431	0.234	0.429	0.260	0.855	0.384	0.466	0.242	0.519	0.295
DE	**0.440**	**0.418**	0.479	0.443	0.574	0.498	0.461	0.432	0.540	0.423	0.568	0.496	0.502	0.446	0.520	0.463	0.565	0.497	0.483	0.447	0.674	0.544
AVG	**0.304**	**0.265**	0.335	0.289	0.412	0.339	0.330	0.282	0.362	0.284	0.412	0.338	0.340	0.290	0.366	0.314	0.533	0.384	0.360	0.303	0.453	0.368

Experiment--长期预测

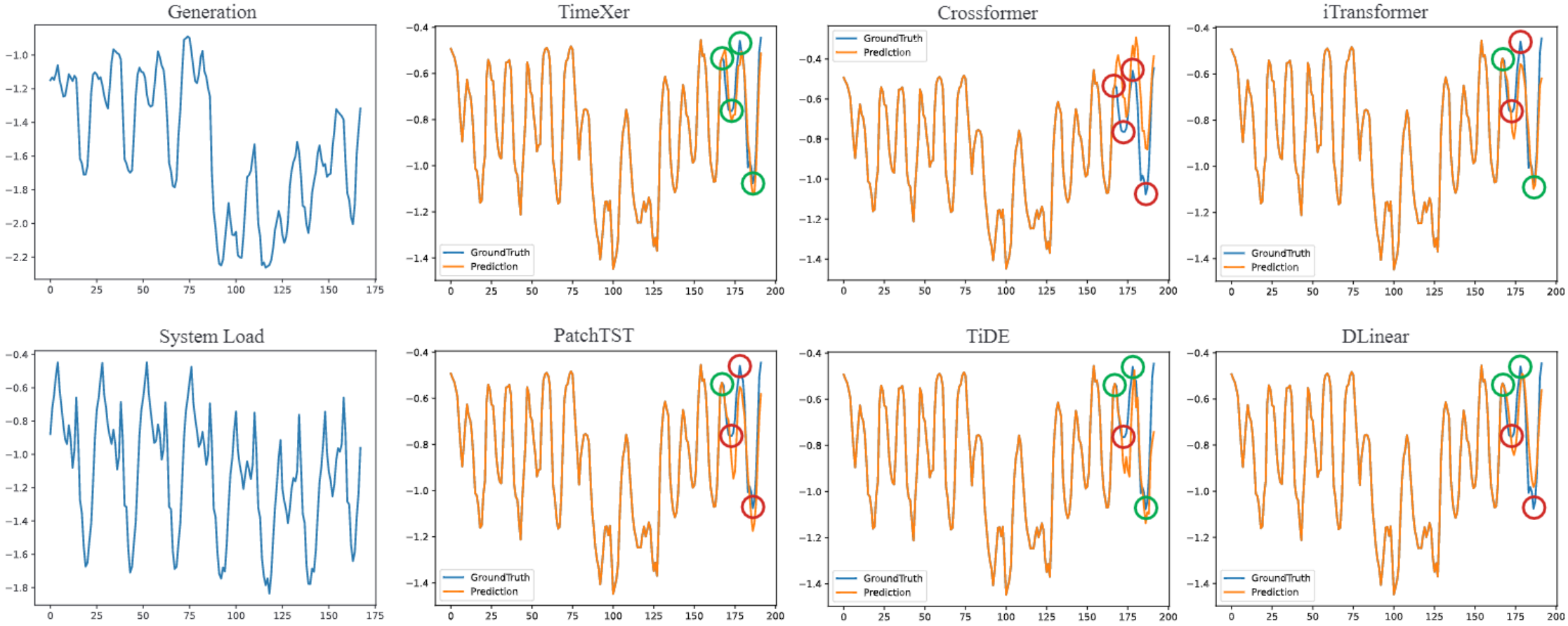
输入长度-96，patch长度-16 预测长度-{96,192,336,720}

Model	TimeXer (Ours)		iTrans. (2023)		RLinear (2023)		PatchTST (2022)		Cross. (2022)		TiDE (2023)		TimesNet (2023a)		DLinear (2023)		SCINet (2022a)		Stationary (2022b)		Auto. (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	**0.336**	**0.414**	0.365	0.442	0.444	0.486	0.394	0.446	0.344	0.412	0.419	0.468	0.410	0.476	0.393	0.457	0.427	0.490	0.372	0.450	0.495	0.528
Weather	0.002	0.031	0.002	0.031	**0.002**	**0.029**	0.002	0.031	0.005	0.055	**0.002**	**0.029**	0.097	0.115	0.006	0.066	0.007	0.071	0.002	0.031	0.006	0.060
ETTh1	**0.074**	**0.210**	0.075	0.211	0.084	0.224	0.078	0.215	0.285	0.447	0.083	0.223	0.076	0.215	0.116	0.259	0.437	0.565	0.110	0.256	0.130	0.282
ETTh2	**0.183**	**0.337**	0.199	0.352	0.205	0.356	0.192	0.345	1.027	0.873	0.205	0.356	0.210	0.362	0.224	0.369	1.155	0.955	0.262	0.405	0.242	0.386
ETTm1	**0.051**	**0.169**	0.053	0.175	0.053	0.173	0.053	0.173	0.411	0.548	0.053	0.173	0.054	0.175	0.066	0.188	0.098	0.241	0.077	0.204	0.085	0.230
ETTm2	**0.116**	**0.252**	0.127	0.267	0.122	0.261	0.120	0.258	0.976	0.769	0.122	0.261	0.129	0.271	0.126	0.263	0.685	0.713	0.207	0.333	0.154	0.305
Traffic	**0.150**	**0.227**	0.161	0.246	0.324	0.412	0.173	0.253	-	-	0.324	0.411	0.171	0.264	0.323	0.404	0.447	0.500	0.361	0.361	0.302	0.353

Experiment--ShowCase

曲线拐点用于评估预测质量。如果预测值保持在 0.05 的范围内，与基本事实相比，我们将此预测视为成功的预测，并在其周围放置一个半径为 0.05 的绿色圆圈。否则，我们将此预测作为超出范围的预测，并将半径为 0.05 的红圈标记为其故障。

序列长度168， 预测长度24



TimeXer更准确的预测，比其他模型更稳健

消融实验--屏蔽外生变量

使用完整的内生序列 对外生序列的掩码从0~99%

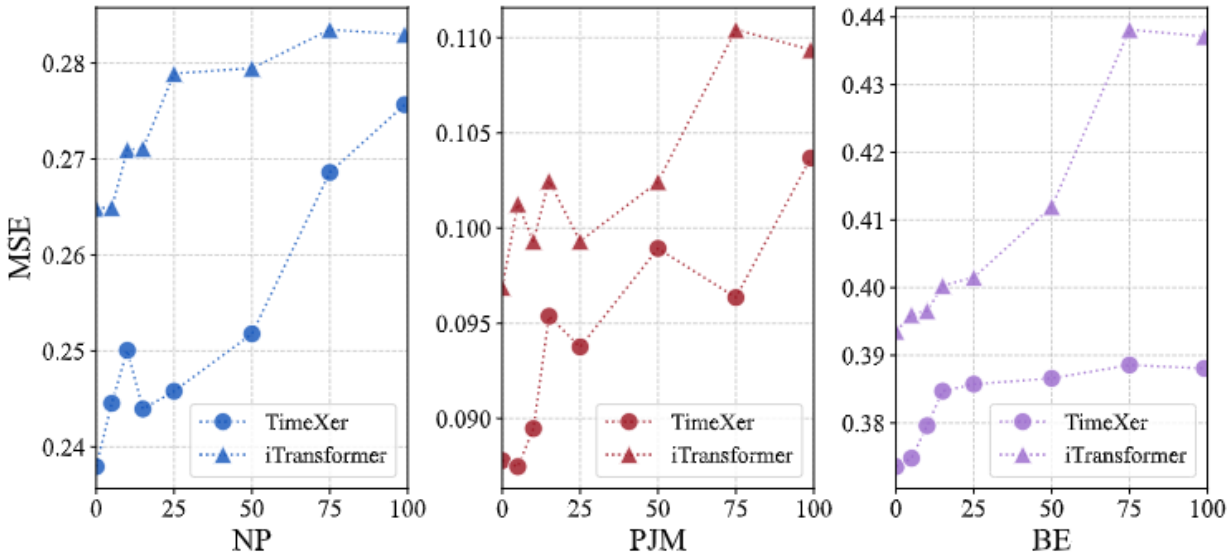


Figure 4. Forecasting performance with the masked exogenous series on three EPF datasets, simulating the missing data scenario.

可以观察到，随着外生序列质量的降低，模型预测性能也会随之降低，但TimeXer仍然表现得比较好，说明TimeXer能够支持低质量的数据场景。

消融实验--嵌入方式的有效性

Table 6. Full results of the ablation study.

Design	Endogenous	Exogenous	Horizon	ETTh2		ETTm2		Traffic	
				MSE	MAE	MSE	MAE	MSE	MAE
Ours	Temporal+Variate	Variate	96	0.130	0.278	0.062	0.180	0.145	0.219
			192	0.179	0.330	0.095	0.229	0.146	0.220
			336	0.209	0.366	0.127	0.270	0.145	0.224
			720	0.215	0.372	0.180	0.330	0.165	0.246
			Avg	**0.183**	**0.337**	**0.116**	**0.252**	**0.150**	**0.227**
Replace	Temporal+Variate	Temporal	96	0.131	0.279	0.067	0.186	0.155	0.234
			192	0.180	0.332	0.100	0.235	0.152	0.231
			336	0.222	0.375	0.135	0.280	0.151	0.233
			720	0.234	0.387	0.185	0.335	0.174	0.258
			Avg	0.192	0.343	0.122	0.259	0.158	0.239
w/o	w/o Variate	Variate	96	0.132	0.279	0.064	0.183	0.153	0.230
			192	0.183	0.335	0.099	0.236	0.152	0.230
			336	0.221	0.375	0.127	0.270	0.151	0.234
			720	0.249	0.398	0.179	0.329	0.175	0.260
			Avg	0.197	0.347	0.117	0.255	0.158	0.239
	w/o Temporal	Variate	96	0.133	0.282	0.063	0.183	0.149	0.226
			192	0.176	0.329	0.097	0.234	0.147	0.224
			336	0.211	0.367	0.129	0.275	0.145	0.230
			720	0.234	0.387	0.181	0.332	0.166	0.251
			Avg	0.188	0.341	0.118	0.256	0.152	0.233

消除任何一种类型的嵌入都会使模型性能下降。

模型分析--变量之间的相关性

交叉注意力机制中的内生变量和外生变量之间的相关性的注意力图

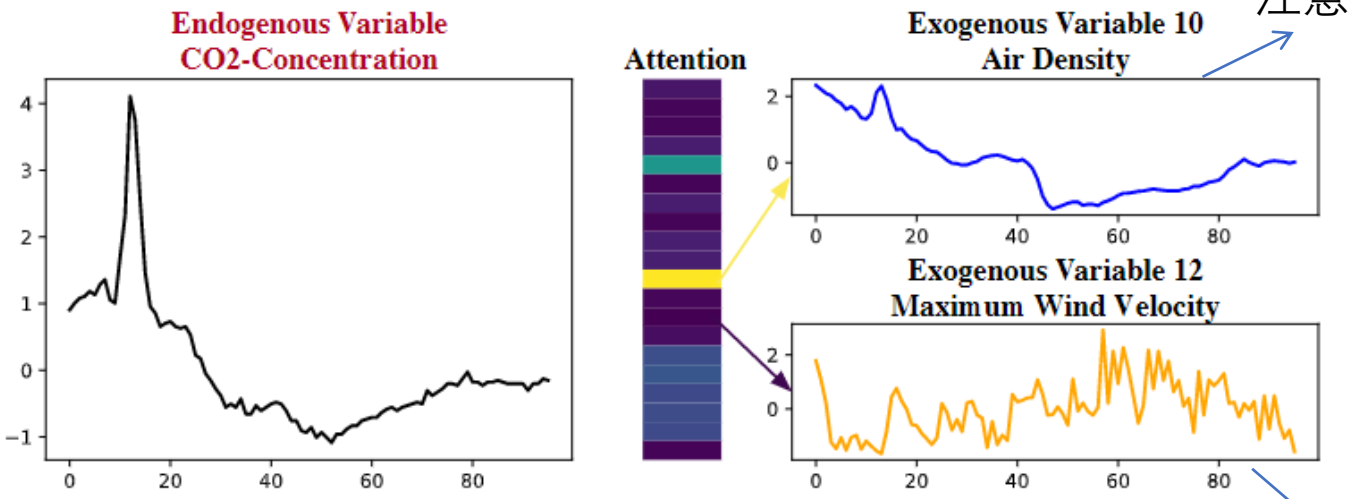


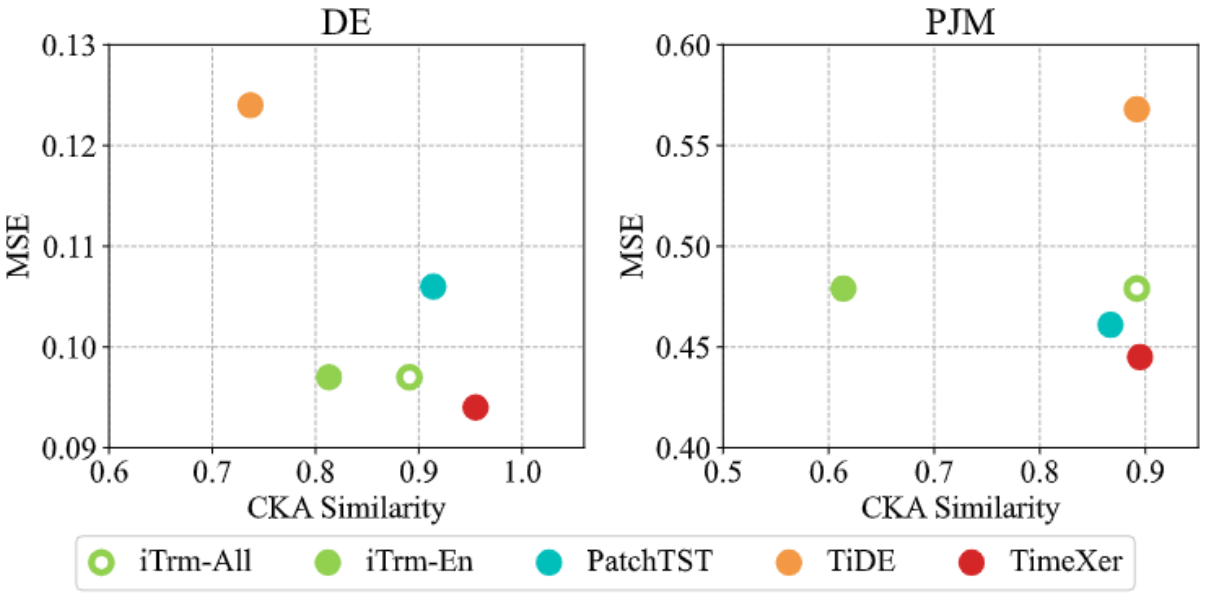
Figure 5. Visualization of learned attention map alongside the endogenous time series and the exogenous time series with highest and lowest attention scores.

使用了交叉注意力机制后，内生变量与外生变量之间相关性的可解释性更强，（有关的变量受到的关注更多，无关的变量受到的关注更少），注意力图更集中。

模型分析--表示分析

CKA（中心核对齐）相似度：

多层模型中，第一层与最后一层表示之间的相似度，越高的相似度代表越好的模型效果



绿色空心点：itransformer用所有序列表示；

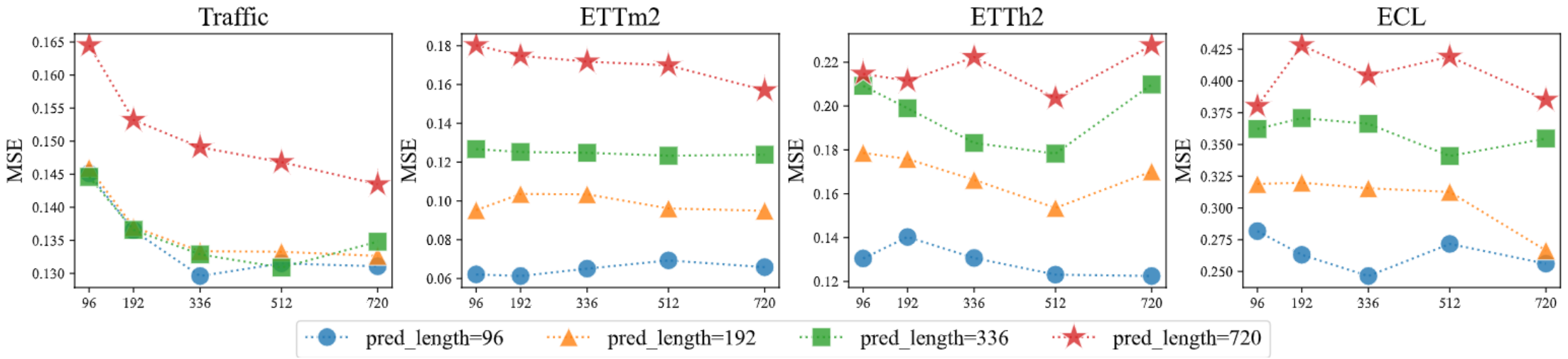
绿色点：itransformer只用内生变量表示

对于MSE，越低越好，CKA越高越好，即图中位于右下角的位置代表模型效果更好。

两类绿色点的结果表明：直接应用多元变量进行预测相较于只用内生变量预测，虽然相似度更好，但效果却没有提升，说明直接使用多元变量会引入一些不必要的噪声，干预预测性能。

模型分析--外生变量的回望窗口

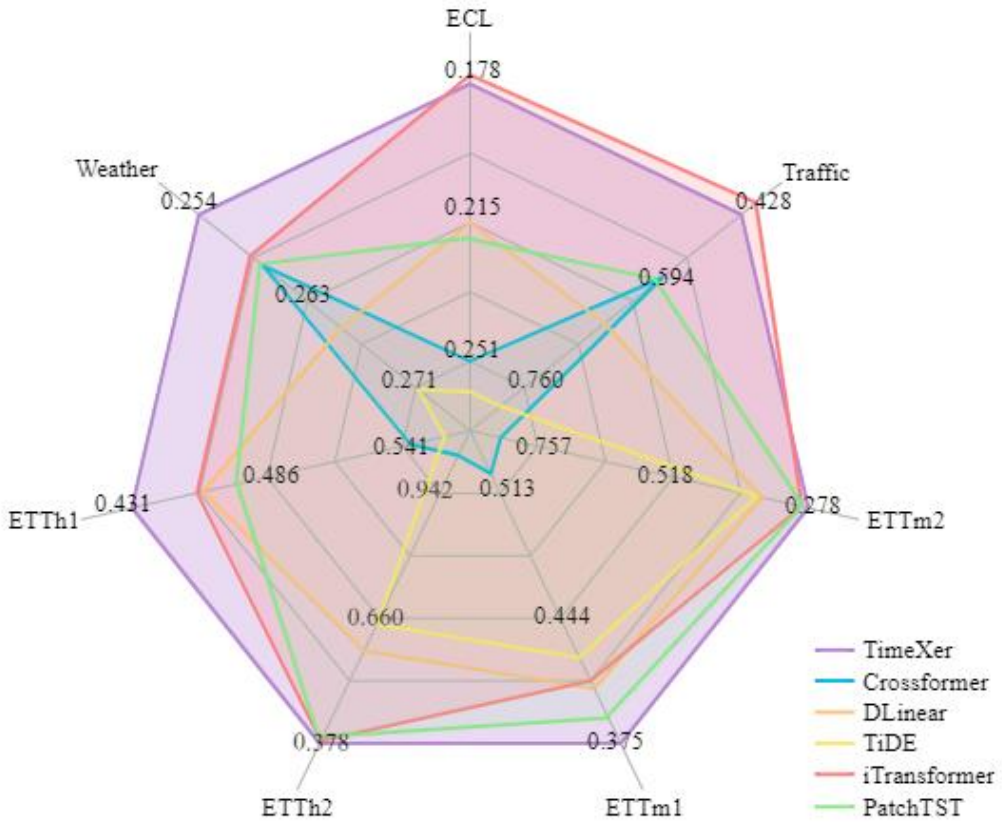
内生变量的长度固定为96，增加外生变量的长度（横轴）



在大多数情况下，预测性能受益于外生回溯长度的增加

模型分析--多变量预测

数据集中的变量视为相互独立的内生变量，每个变量考虑所有其他变量作为外生变量



可以观察到，TimeXer在基线模型的多变量预测任务中表现出了具有竞争力的性能，突出了其有效性和通用性

西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions