## 相似性分组引导神经网络在海事时间序列预测中的建模

摘要

可靠、准确的时间序列预测对于航运业的经济投资、交通规划、港口规划、设计等都起着至关重要的作用。海事时间序列的动态增长具有复杂、非线性和非静止性质。为了保证高质量的预测性能，我们首先提出采用经验模态分解和集成经验模态分解方法将原始使劲啊序列分解为高频和低频部分。低频部分可以直接用简单的传统神经网络预测。由于高频部分没有较强的数学规律所以很难去预测。为了利用高频分量内在的自相似性，将这些分量划分为几个连续的小(重叠)段。然后选择相似度高的分组片段形成更适合传统神经网络方法的训练数据集。这种重组策略有助于提高高频部分的预测精度。对预测的高频分量和低频分量进行积分，得到最终的预测结果。我们提出的三步预测框架受益于时间序列的分解和类似的片段分组。对港口货物吞吐量和船舶交通流的实验表明，该方法在预测精度和鲁棒性方面都具有较好的性能

关键词

数据预测，自然网络，最小相似度组，经验模式分析，动态时间规划

I. 介绍

A. 背景及相关工作

海上时间序列的动态增长(如港口货物吞吐量和船舶交通流)有时具有主要的复杂、非线性和非平稳特性。其可靠、准确的预测对航运业的经济投资、交通规划、港口规划设计等都起着重要作用。在文献中，最近致力于有效地预测不同类型的时间序列而做努力。然而，使用传统数学方法仍是比较困难的建模和预测如非静止时序，如模糊集合论、卡尔曼滤波、贝叶斯模型、混合框架、自回归积分滑动平均模型和他们的变种。

站在机器学习的观点去看，有不同的方法被引入到时间序列预测中。最常见的两种方法是传统的神经网络和最新的深度学习。它们都是有监督的学习方法，被训练去学习几个输入特征和输出值之间的映射函数，输出值被表示为要预测的目标。近年来的研究表明，深度学习在城市交通流等时间序列预测中具有广阔的应用前景，如城市交通流、晶片编号的周期时间、风速、金融市场等。然而，深度学习的预测基础是高度依赖于训练数据集的数量和多样性。如果训练的样本集合没有包括高度相似具有特点的显示在时间序列去越策，会造成困难的产生满意的结果。高计算复杂度也限制了基于深度学习的预测方法在实际应用中的可部署性。因此，为了使预测更容易、更灵活，本文主要研究神经网络和它在非平稳时间序列预测中的扩展

神经网络是否有能力学习不同变量之间潜在的复杂关系。传统神经网络方法，例如：小波神经网络（WNN），模糊神经网络（FNN），埃尔曼神经网络（ENN）,误差反向传播神经网络（BPNN），回归神经网络（GRNN）等。在过去的很多年里吸引了许多关注，海事时间序列的发展往往受到多种复杂因素的影响。由于神经网络方法有其自身的缺点，仍然难以产生满意的预测结果。例如，反向传播神经网路在训练过程中会陷入局部极小问题；回归神经网络经常遭受高计算时间的问题。

为了进一步提高预测性能，经验模态分解(EMD)和神经网络相结合，形成了一个两步预测框架。例如，组合版本有EMD+WNN，EMD+FNN，EMD+ENN，EMD+BPNN，EMD+GRNN在时间序列预测方面表现良好。这是因为EMD能够将原始的非平稳时间序列分解成一系列独立的、近似周期的分量，这些分量可以接受性能良好的希尔伯特变换。每个分量与一个本征模态函数(IMF)相关，可以揭示时间序列的隐藏模式和趋势。特别是，频率最低的IMF分量通常表示原始时间序列的趋势或平均值。因此，这种分解策略能够有效地帮助开发预测方法。众所周知，传统的经验模态分解容易出现模态混合问题。因此，这种分解策略能够有效地帮助开发预测方法。众所周知，传统的经验模态分解容易出现模态混合问题。为了进一步提高数据分解性能，Wu和Huang在2009年提出了一种新的噪声辅助方法——集成EMD (ensemble EMD, EEMD)。因此，EEMD+WNN、EEMD+ENN、EEMD+BPNN、EEMD+GRNN[36]在时间序列数据预测领域也得到了广泛的关注。对于传统的基于EMD和EEMD为基础的预测方法，分别使用传统的神经网络方法(如WNN、FNN、ENN、BPNN和GRNN)对高频和低频分量进行预测。将预测的高频和低频分量进行聚合，得到最终结果。此外，最近还将EMD和EEMD与新开发的深度学习方法相结合，进一步提高了预测结果。

B 动机和贡献

一般认为，原始的时间序列通常是由几个高频率和低频率组成的时间序列的不同特征对应的分量。在目前流行的两步预测框架中，使用WNN、FNN、ENN、BPNN、和GRNN可以直接准确预测变化稳定的低频分量。然而，由于高频分量具有不明显的数学规律性，使得高频分量的预测变得困难。在实际应用中，基于神经网络的预测结果可能会导致非最优结果。我们发现，高频分量的自相似性可以实现稳健而准确的预测（这意味着每一个高频分量理论上包含几个小的(重叠的)片段，这些片段具有相似的几何形状）。在我们提出的三步预测框架中，高频分量首先被分割成几个连续的小(重叠)段。利用动态时间调整算法(DTW)方法可以有效地测量任意两个不同线段之间的几何相似性。与基于标准化欧几里德距离(SED)的相似度度量相比，DTW能够找到形状相似的点，从而提高了不同路段之间的相似度度量。然后选择具有高DTW相似性的分组片段作为传统神经网络方法的训练数据集。它能够提高对高频分量的预测。将所有的预测分量组合起来作为我们所提出方法的最终预测。

综上所述，结合目前的研究工作，我们的主要贡献主要有以下三个方面：

1）针对海事行业的非平稳时间序列预测，提出了统一的三步预测框架(EMD+DTW+NN和EEMD+DTW+EE)。基于EMD或EEMD的属性分解和基于DTW的相似度分组可以获得满意的预测性能。

2）将非平稳时间序列分解为高频和低频分量。在提高预测精度方面，考虑了高频分量内在的自相似性。利用传统的神经网络方法直接对低频分量进行预测。

3）港口货物吞吐量和船舶交通流的综合实验表明，所提出的三步预测框架在预测精度和鲁棒性方面具有较好的性能。

我们提出的预测方法的主要优点是充分利用了基于EMD或EEMD的属性分解和基于DTW的相似度分组。因此，在不同的应用中可以有效地提高预测精度和鲁棒性。

C 组织

本文的其余部分组织如下。提出了基于分解的预测框架在第二部分。第三节专门介绍港口货物吞吐量和船舶交通流量预测的几个数值实验。本文通过总结我们在第四部分的主要贡献来总结。

II。相似分组引导神经网络建模

本节将详细说明我们所提出的三步预测框架，如图.1所示，在第一步中，可以采用EMD或EEMD将原始的非平稳时间序列分解为高频和低频分量。第二步，利用传统的神经网络方法直接对低频分量进行有效预测。相比之下，高频分量首先被分成几个连续的小(重叠)段。然后将具有高于基于DTW相似度的多个片段进行分组，形成新的训练数据集，用于传统的神经网络方法，以提高预测精度。在最后一步，将预测的高频分量和低频分量进行组合，形成最终的预测结果。

A 基于EMD的时间序列分解

在文献中，EMD被用来将非平稳信号(即时间序列)分解成一个小数目的不同尺度的数据序列。每个数列都可以指定为一个本征模态函数(IMF)，从数学的观点来看，本征模态函数是独立的和近似周期的。DTW的实现必须满足以下两个条件：

1）整个信号的过零点和极值点(局部极大值和局部极小值)的个数应该相等或相差不超过1。

2）在信号中的任何一点，由局部极小值和局部极大值表示的上下包络的平均值应该等于零。本质上，经验模态分解高度依赖于原始信号的局部特性，如局部极小值、局部极大值和过零。将原始非平稳信号分解为imf通常使用一种称为筛选算法的迭代过程。特别是，它计算了从细到粗的每个尺度上的imf。有关EMD的更多细节，我们建议感兴趣的读者查看文献43和其中的参考资料。设X(t)表示原始的非平稳时间序列，可以将其分解为有限个imf为1≤t≤T。其中，N为总IMF个数，Ci(t)为第i个IMF, RN (t)为原始数据X(t)的均值趋势的残差信号。在实践中，RN (t)也可以看作是一个IMF，记作CN+1(t)。例如，（1）可以写成怕【3左下】在这个分解中，第一个IMF与波动最快的部分有关X(t)，同时最后一个IMF对应于波动最慢的部分

B 基于EEMD的时间序列分解

在第二-A节中介绍的传统经验模态分解很容易受到模态混合问题的影响。这一问题容易导致时频分布出现严重的混叠，降低分解精度。为了提高分解性能，Wu和Huang提出了一种新的噪声辅助方法——EEMD，该方法与EMD中的迭代筛选过程相同。利用传统的经验模态分解方法将时间序列数据反复分解成不同的本征函数。特别地，EEMD中的IMF分量被定义为轨迹均值的集合。每次跟踪由时间序列数据的分解结果加上均匀分布的有限振幅白噪声组成。理论和实践结果都表明，在EMD方法中加入的噪声对数据分析有一定的帮助。EEMD的具体步骤如下：

1）在原始时间序列数据中加入均匀分布的白噪声序列。

2）将加入白噪声的时间序列数据通过章节二-A所示的EMD分解成不同的imf。

3）用不同的白噪声重复步骤(1)和(2)，得到相应的IMF分量。集合数本质上与重复过程的次数有关。

4）计算集合本征函数的均值作为最终分解结果。

对于基于EMD的时间序列分解和基于eemd的时间序列分解，本工作将imf看作是高频和低频分量的组合。在实际应用中，利用传统的神经网络方法可以很容易地直接预测出低频分量。而高频分量的数学规律性较弱，难以预测。

C DTW-BASED相似性分组

DTW是一种基于动态规划方法的度量两个时间序列之间距离(或相似性)的有效且流行的方法。在语音识别、数据聚类、特征提取等方面得到了广泛的应用。两个时间序列之间的相似性与几何距离成反比。DTW算法的基本原理是比较两个时间序列，通过计算两个时间序列之间的最小距离来度量它们的相似性。

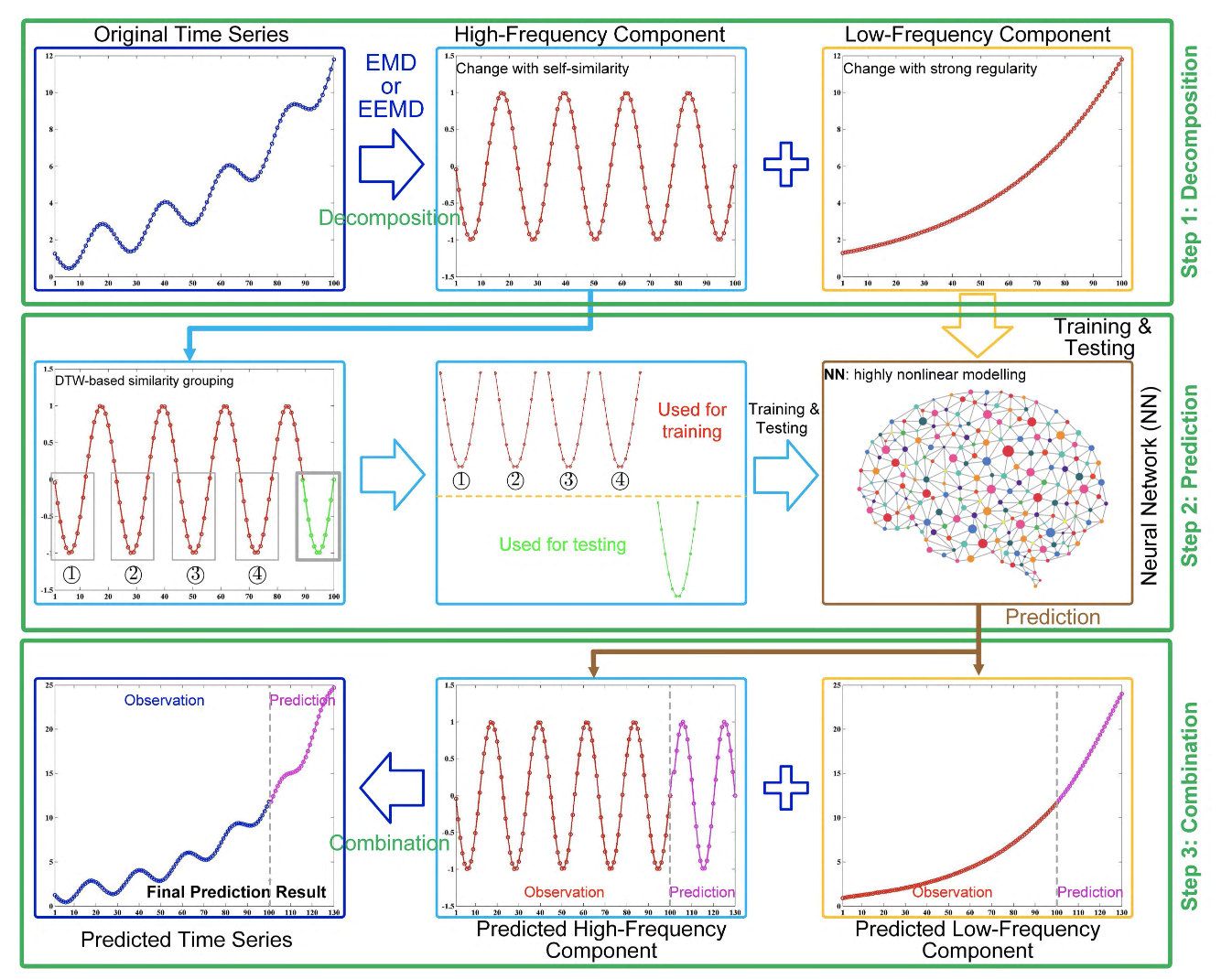


图1，提出的EMD+DTW+NN和EEMD+DTW+NN三步预测框架的流程图。为了简化，将原始时间序列分解为两个分量。在实际应用中，为了保证高精度的预测，需要生成两个以上的分量。

设Y = {y1,y2，…，ym}和Z = {z1,z2，…，zn}表示两个时间序列。在计算距离之前，我们首先创建一个m×n的矩阵其中(ith,jth) 元素代表距离d (yi,zj)表示两点yi和zj之间的。为了增强计算鲁棒性，d (yi,zj)对应加权欧氏距离。因此，基于dtw的相似度对噪声或离群值不敏感。如图1所示，很容易获得m = n。这两个序列Y和Z之间的最佳匹配对应于一个序列与另一个序列对齐的最短距离路径。因此，可以递归地计算出最优的变化碎片。DTW (Y,Z) =γ (i,j), (2)

其中最小累积距离γ (i,j)由：

γ (i,j) =d (yi,zj) +min {γ (i −1,j −1),γ (i −1,j),γ (i,j −1)}. （3）

为了更好地理解，图2显示了DTW距离和欧几里德距离之间的区别。与广泛应用的欧氏距离相比，DTW在提取几何形状相似点时具有更好的性能，提高了距离(相似度)度量的准确性。值得注意的是高频分量(i.e., IMFs)由于它们具有高波动性的特性，很难预测。在这项工作中，高频imf将被划分为一系列重叠的片段。利用DTW距离测量相似度高的线段合并到同一组中。聚集的群体被利用来形成为传统的神经网络方法提供更合适的训练数据集，以提高预测性能。

D 三步预测框架

所提出的非平稳时间序列预测的三步框架如图1所示。EMD、EEMD和DTW的基本原理已在第II-A-II-C节中介绍。我们的EMD+DTW+NN和EEMD+DTW+NN的具体计算步骤如下

1）第一步：分解

EMD或EEMD方法首先将原始的非平稳时间序列X分解为高频IMFsCh1≤p≤P和低频IMFsCl1≤p≤P在P+Q=N+1时。尽管Ch1≤p≤P可以提取出X的主要结构和特征，但由于Ch1≤p≤P的数学规律性较弱，仍然难以预测。相比之下，Cl1≤p≤P更容易预测，因为它表明了长周期特征。值得注意的是，最后一个IMF ClQ一般代表了X的基本趋势，可以保证其预测的高精度。为了简单起见，在本文的其余部分，我们倾向于分别采用Chp和Clq，而不是Ch1≤p≤P和Cl1≤p≤Q。

2) 第二步：预测

利用传统的神经网络方法，如WNN、FNN、ENN、BPNN和GRNN等，直接预测低频IMFsClq。这五种神经网络方法背后的数学基础在目前的研究工作中已经被广泛讨论。为了尽可能的简洁和更好的理解，本文只介绍相应的建筑图表，如图3所示。由于高频IMFsChp具有高波动性的特性，因此很难对其进行准确的预测。为了克服这一缺点，我们提出利用Chp内部的自相似性。特别地，每个Chp首先被划分为一系列重叠的部分，如Chp={{Chp (1), Chp (2), ···· · ·,Chp (L)}, ···· · · , {Chp (i), Chp (i+1), ···· · ·Chp (L+i-1)}, ·· · · , {Chp (T-L+1), Chp (T-L+2), ···· · ·,Chp (T)}} ，其中L表示每个线段的长度。如何为神经网络方法构建合适的训练数据集对最终的预测性能有重要影响。以Chp (T+1)的预测为例，我们认为Ch,T-L+1p={ Chp (T-L+1), Chp (T-L+2), ···· · ·,Chp (T)} 作为输入参考段。然后利用DTW方法计算Ch,T-L+1与Chp中其他段的相似度，如图2所示。DTW具有生成不同段之间相似度稳健3估计的能力。

在本工作中，我们提出将Chp中相似度较高的片段自动分组，形成更合适的训练数据集，用于基于神经网络的高频分量Chp(T+1)预测。与此策略类似，可以实现对Chp(T+·)的准确预测。值得注意的是，在Chp中，如果某些片段与参考片段的相似性程度较低，则不适合促进预测。在该框架内，只提取相似性高的分组片段，提高了预测的准确性和鲁棒性。

3）第三步：结合

利用EMD+DTW+NN和EEMD+DTW+NN对高频IMFsCh1≤p≤Pp进行了预测，提高了预测精度，而低频IMFCl1≤p≤Q则直接用传统的神经网络方法进行预测来缩短计算时间。值得注意的是，EMD+DTW+NN和EEMD+DTW+NN可以进一步提高低频IMF的预测精度，但代价是计算量大大增加。在本工作中，最终的预测结果是通过对预测的高频和低频IMF进行积分得到的。

  (4)

对于1≤≤ LT, LT是数值实验中需要预测的时间序列数据点的长度。通过提出的两种三步预测框架，能够产生更准确、更稳定的预测性能。

III．实验结果与讨论

1. 港口货物吞吐量(pct)

在这项工作中，我们选择了现实的PCT时间序列，从中国18个主要港口从1996年到2016年。PCT时间序列的增长趋势如图4所示。结果表明，由于地理位置、货物种类、港口规模、服务水平、管理水平等因素的影响，不同港口的增长趋势具有不同的性质。这意味着所提出的预测方法应具有通用性和鲁棒性，以提高预测结果。

1. 船舶交通流(VTF)

我们建议在武汉长江大桥采用原有的VTF时间序列来评价中期预测性能。利用动态自动识别系统（AIS）[AIS是一种用于船舶的自动跟踪系统，能够实时提供船舶的位置(经纬度)、航向和速度等。]可以生成原始的VTF数据。特别是每隔3小时提取16天的VTF数据，以辅助中期预测。原始的VTF数据见第三- d节的图10。据我们所知，以往的工作主要集中在长期VTF预测上。与以往的研究相比，我们的VTF时间序列具有更高的时间分辨率。中短期预测结果有利于海事行业精细化管理和提高交通安全水平。

B．与其他预测方法的比较

本文提出的EMD+DTW+NN和EEMD+DTW+NN的三步预测框架，将主要与传统的预测方法(即NN和EMD+NN)进行如下比较

神经网络:在不失一般性的前提下，引入了广泛应用的神经网络方法，如小波神经网络、模糊神经网络、新神经网络、BPNN和GRNN，直接预测航运业的非平稳时间序列。在这5种常用的神经网络方法的基础上，开发了EMD+NN和EMD+DTW+NN框架，进一步提高了预测性能。

EMD+NN: EMD+NN本质上是一种结合了EMD和传统的两步预测框架神经网络方法。和文章，[24]，[46]一样，我们先用EMD方法将原始的非平稳时间序列分解为几个高频和低频分量。然后利用神经网络方法直接预测高频和低频分量。将预测的高频分量和低频分量相结合，得到最终的预测结果。从理论上讲，EMD+NN与单一神经网络相比具有更强的鲁棒性。

EMD+DTW+NN:这种三步预测框架可以被认为是EMD+NN的扩展，充分利用了高频分量的自相似特征。在此框架下，利用常用的神经网络方法直接对低频分量进行预测。该方法将高频分量划分为若干小段，通过基于DTW的相似度度量对小段进行分组。对分组的片段进行选择，形成更合适的训练数据集供神经网络使用，提高神经网络的预测精度。两者都预测了高频和低频补偿。

EEMD+DTW+NN:该预测框架本质上是EMD+DTW+NN的扩展，在时间序列分解时用EEMD代替EMD。EEMD是一种噪声辅助的数据分析方法与简单的EMD相比，EEMD生成的分解结果更鲁棒、更稳定。理论上基于EEMD的三步计算框架能够产生较好的预测结果。

为了定量分析预测结果，选择相对误差(RE)作为准确性的度量。相应的数学公式定义如下：

 （5）

其中y和分别为实际时间序列和预测时间序列。为了进一步评估预测的稳健性，在我们的数值实验中，每种预测方法将运行10次来生成正则表达式的均值和RE的标准差版本。

C．港口货物吞吐量预测实验

1）量化绩效评估

为了实现实验，我们使用图4中提取的1996年至2015年中国18个主要港口的PCT时间序列来训练我们的三步预测框架。利用2016年所有这些港口的预测数据对预测精度进行了评估。可以发现，秦皇岛港PCT时间序列的增长趋势更加复杂。从数学的角度来看,由于明显的规律变化,更容准确预测时间序列预测PCT在其他港口通过直接使用传统技术,如卡尔曼滤波,非参数回归,小波滤波,神经网络方法等。因此，以秦皇岛港为例，我们提出将PCT时间序列分解为4个不同的分量，如图5所示。第一个通常与高频IMF相关，于是很难预测。其他3个分量对应低频的IMF。此工作中的其他端口也将执行类似的分解方式。

在我们的数值实验中，我们经验地选择公式(4)中的P =1和Q = 3，因为大量的实验结果表明，这种选择能够产生令人满意的性能。我们提出的框架EMD+DTW+NN、EEMD+DTW+NN和其他竞争技术(即NN和EMD+NN)的定量结果如图6所示。可以看出直接使用神经网络的预测精度最低。而EMD与神经网络相结合可以提高预测性能。对于秦皇岛港口PCT数据的趋势更加复杂，我们提出的方法EMD+DTW+NN和EEMD+DTW+NN的预测结果优于NN和EMD+NN。这是因为基于EMD和EEMD的PCT分解产生相对平稳的分量，可以很容易地使用NN方法。DTW能够稳健地度量被分解构件内部不同小段之间的相似度。通过选择相似度高的片段对神经网络方法的训练数据集进行分组，提高神经网络方法的预测性能。因此，我们所提出的两种三步框架可以显著优于其他竞争方法。相应的预测结果如图6所示，更接近实际PCT数据的虚线。由于EEMD本质上是EMD的改进版本，所以EEMD+DTW+NN产生的预测结果更接近实际数据。

图7和图8所示的预测精度和鲁棒性的统计结果进一步证实了我们所提出的框架的优越性能。RE的均值【正则表达式的标准差本质上与预测性能的鲁棒性有关。特别是在实际应用中，标准差较低，表明预测结果是稳健的。】结果在图7证明EMD+DTW+NN和EEMD+DTW+NN在所有条件下都能产生最准确的预测。而EMD+NN则限制了预测性能的进一步提高。由于PCT时间序列具有较弱的数学规律性，在大多数情况下，单神经网络方法的预测精度最低。图8直观地展示了RE3用于评估预测鲁棒性的标准差(std)。可以发现EMD+DTW+NN和EEMD+DTW+NN都能够产生最稳健的预测。这种良好的性能主要得益于基于EMD或eemd的属性分解和基于dtw的相似度分组。

1. 预测结果与讨论

显然，在所有的神经网络方法中，神经网络具有最高的预测精度和鲁棒性，如图6-8所示。因此，本小节将我们的三步计算框架与2017 - 2026年PCT预测的其他竞争方法(即ENN和EMD+ENN)进行比较。预测结果显示在图九，说明了不同预测框架之间的比较。不难发现，我们的EMD+DTW+ENN和EEMD+DTW+ENN对PCT时间序列的预测具有鲁棒性。从图4的原始数据可以看出，从1996年到2016年，大部分港口的PCT呈稳步上升的趋势。从理论上看，2017年至2026年的预测结果也应该会稳步增长。从图9可以看出，我们所提出的框架生成的预测PCT时间序列更接近理论结果。新神经网络和EMD+新神经网络都容易出现计算不稳定的问题，导致预测性能低下。通过充分利用属性分解和相似分组的优点，我们提出的三步计算方法能够充分把握PCT时间序列的变化趋势，提高预测精度。特别是，EMD和EEMD可以将原始PCT时间序列分解为有限个分量的和。各组分的波动性减小，提高了基于神经网络的预测精度。实际应用表明，该方法能够准确、稳健地预测PCT时间序列。