福州大学

# 本科生毕业设计（论文）

题　　目：基于聚类算法和长序列循环神经网络的旅行时间预测

姓　　名：刘奕薇

学　　号：031703120

学　　院：数学与计算机科学学院

专　　业：信息安全

年　　级：2017 级

校内指导教师：_____（签名）

校外指导教师：_____（签名）

2021 年 5 月 18 日

# 基于聚类算法和长序列循环神经网络的旅行时间预测

## 摘要

随着人们的出行频率与日递增，大众期望可以更准确地预估在路程中所需花费的时间，从而提高出行成本和体验。车辆经过一个路段所花费的时间被定义为该路段的旅行时间。而旅行时间的预测在本质上是时间序列的回归问题。但是与传统的回归问题不同，在寻找旅行时间变化趋势的最佳拟合曲线时，不仅需要考虑目标路段在历史时段的旅行时间，也需要参考与之相邻的路段的交通状况。因此，如何有效地获取时间因子和空间因子对于序列变化的影响是旅行时间预测研究中的一个重大挑战。而近年来，机器学习算法和深度学习模型的发展帮助研究人员有效地利用交通数据的时空关联性来提高旅行时间的预测准确率。与此同时，有许多研究证明，在回归分析的研究领域，对于具有不同变化趋势的子数据集分别建立模型进行预测的方法有助于提高总体目标的回归准确率。那么，如何利用分段回归方法的优点来提高旅行时间的预测准确率是一个亟待探讨的问题。因此，为了有效利用路段的时空因子以及分段回归方法来降低旅行时间的预测误差，本文提出基于空间特征层次聚类算法(Spatial Feature-based Hierarchical Clustering, SFHC)和深度多输入门控循环单元网络模型（Deep Multi-input Gated Recurrent Unit, DM-GRU）的旅行时间预测方法。该方法首先利用 SFHC 算法把交通网络分割成多个路段群，群中的所有路段具有相似的空间特征。之后，在每个路段群中建立一个利用时空依赖和外部信息的 DM-GRU 模型来预测该群中每个路段的旅行时间。在福建省漳州市某交通区域上展开的相关实验证明，本研究提出的方法比基准模型取得了更小的预测误差。此外，本研究对比并分析了 SFHC 算法和其他三种经典的聚类方法在挖掘路段之间的空间关联性以及提升旅行时间分段预测效果两方面的差异。对照实验的结果表明，SFHC 算法在分段预测中取得了比其他聚类算法更小的误差。因此，本研究提出的方法是一个有效的旅行时间预测方式，而该研究成果可以应用在路线规划、智能交通等领域。

**关键词：旅行时间预测，分段回归分析，层次聚类，门控循环单元**

# Travel Time Forecasting based on Clustering Algorithm and Recurrent Neural Network

## Abstract

As traffic demand boosts with increasing urban level, people are expecting to get a more accurate prediction on the time they need to spend when commuting, as a result, they can reduce transportation cost and have a more pleasant trip on the road. By definition, the amount of time spent by vehicles passing a road segment is usually called the travel time of this segment, while forecasting travel time is a time series regression problem. But different from ordinary regression problems, travel time of the future is usually affected by data from previous time steps and traffic variety of adjacent segments. Then, how to make use of temporal and spatial factors in traffic records to improve prediction accuracy is a challenge for researchers. Thanks to the development of machine learning algorithms and deep learning models in recent years, researchers are able to utilize these special factors of traffic data with these techniques. At the same time, it has been proved by much research that, building individual regression models for sub-dataset with different trends contributes to improving overall regression accuracy. But how to predict travel time in this piecewise manner is an interesting topic. In order to facilitate temporal and spatial features of travel time series and adopt the method of segmented regression analysis, this paper proposed a unique travel time forecasting method combining a deep multi-input gated recurrent unit model (DM-GRU) with the spatial feature-based hierarchical clustering algorithm (SFHC). First, SFHC partitions traffic networks into clusters, where segments have similar spatial features with each other. Then, a DM-GRU model, which exploits temporal and spatial dependencies and external factors, is constructed in each cluster for predicting travel time of every segment inside. Experiments carried in an urban area in Zhangzhou, Fujian, showed that the proposed method achieved lower regression errors than baseline models. In addition, we compared the effects of SFHC on segmented travel time regression analysis with three other classic clustering algorithms, as well as analyzing spatial correlation between segments. According to the results of comparison experiments, SFHC outperformed other clustering methods in piecewise travel time prediction. Hence, the proposed method in this paper is effective for forecasting travel time of road segments, and it is a promising technique for route planning and intelligent transportation systems.

**Key words: travel time prediction, segmented regression analysis, hierarchical clustering, gated recurrent unit**

# 目录

# 第 1 章 绪论

## 1.1 引言

近年来，各大城市中日趋严重的交通拥堵问题，给人们的出行体验和成本带来极大的负面影响。因此，如何合理规划出行路线，以便节省通行时间和成本的问题逐渐受到人们的重视。在路线规划时，能否获得未来时刻精确的交通状况是一个至关重要的因素，而交通状况的内容一般包含道路的旅行时间、交通流量以及速度。其中，道路的旅行时间被定义为车辆从此路的起始点到其终止点所需花费的时间。这一概念直观地反映出某区域内的交通状况，而人们在规划路线时，也最先考虑到这个因素。因此，如何准确预测旅行时间是一个值得关注的研究方向。

## 1.2 研究背景

从本质上来说，旅行时间的预测属于时间序列的回归问题。针对这一问题的研究，从上世纪末到近五年，依次经历了基于统计学的方法，传统机器学习算法和深度学习模型三个阶段的发展[1]。在 1996 年，基于统计原理的自回归积分滑动平均模型就已被用于预测时间序列[2]。然而该模型要求其研究对象的均值、方差等统计特征为常量。在这种情况下，这些统计性质才可以被用于预测未来时刻的情况。可是现实中未来时间点的交通状态往往由多种因素决定。比如，路段的旅行时间可能会由周围邻近路段的交通状况影响；出行需求的差异导致道路在工作日和节假日的交通状态不同。而在随后的研究中，人们开始使用机器学习算法解决回归问题。比如，贝叶斯网络[3]和隐藏马可夫模型[4]用于建立线性回归模型，而支持向量回归(SVR)用于解决非线性的回归问题[5]。这些方法往往是基于输入因子之间为相互独立的假设。而当输入因子之间存在依赖性时，使用这些回归方法会产生较大的预测误差。因此，研究人员开始利用神经网络获取不同变量之间的影响，并将深度学习模型用于解决交通领域内复杂的回归问题。由于长短期记忆单元(LSTM) 和门控循环单元（GRU）可以解决长序列数据训练时的梯度消失和梯度爆炸的问题，以及卷积神经网络（CNN）可以捕获研究对象不同层次的特征，这两类神经网络及其相关变体被广泛应用到交通方面的回归模型研究中。比如，LSTM 用于预测车辆行程速度[6][7]；堆叠 LSTM 网络用于一次性预估多条路段的交通状态[8]；加入注意力机制的 LSTM 网

络用于捕捉时间序列周期性迁移特征[9]；建立多层 CNN 网络以获取多尺度的行程状况[10]；将交通网络的时空信息矩阵转化为图片，并送入 CNN 网络进行预测未来旅行时间[11]; 使用深度 CNN 网络的优化版本-残差神经网络（Resnet）来预测大规模的交通流量[12]。卷积长短期记忆单元(ConvLSTM)网络是通过在 LSTM 的输入结构中加入卷积核而产生的[13]。相比于 CNN 或 LSTM，由于 ConvLSTM 结合了二者的特征，在交通状况的预测研究中，它已被用于捕捉交通数据在时空维度上的依赖[14]以及其他因素的影响[15]。

与此同时，研究人员也在关注使用聚类算法将单一而复杂的交通状况回归模型划分为多个子模块进行分析的效果。在回归分析的研究领域，分段回归被视为一种解决因数据集包含多种变化趋势而无法获得准确的回归分析结果的有效方法。相比于单一的回归模型，在不同模式的子数据集中建立相应的模型，可以更好地拟合真实的回归曲线[16]。还有研究表明，相对于使用单一模型解决回归问题，分段建模可以大幅减少总的计算时间和内存空间[17]。同时，由于聚类算法常被用于在未带有标签的数据集中挖掘具有相似特征的部分，所以在分段回归分析中，它常被用于划分出具有不同趋势的子数据集。在交通状况的回归分析领域，有研究通过谱聚类[17]、模糊均值聚类[18][19][20]、K 均值[21][22]、层次聚类 [23][24][25] 这四种算法区分出具有不同交通模式的日期或时间段，从而在每个模式中建立相应的回归模型。比如，早晚高峰和非拥堵的时间段、工作日和节假日交通状况一般都有比较明显的区别。因此，以分段的方式进行回归分析，可以捕捉到更详细的交通情况变化趋势，从而在总体上取得比单一模型更好的预测结果[26]。

但是，此类分段建模的方法只关注到交通数据在时间维度上的特征，却忽略了数据的空间特征。在实际生活中，处于相同城市功能区内的路段一般具有相似的旅行时间变化趋势，比如商务区的路段在上午和下午皆有通行高峰，而公园附近的路段在一天中的旅行时间基本是平稳的；上下游节点的交通状况往往会经过一段时间后影响到目标路段。因此，在路段的旅行时间预测问题中，路段自身的交通状况变化趋势和相邻路段的交通情况也需要被考虑到，而本研究将这两种影响因素称为路段的空间特征。从空间特征的角度进行分段预测旅行时间也是一个有意义的研究方向。

虽然已有研究人员尝试从空间特征的角度，对路段的旅行时间进行分段回归分析[27]。但是，其方法是基于传统的层次聚类算法，仅能选出具有相似旅行时间变化趋势的路段群，之后还需要人工处理聚类结果，从而保证路段群中的各路段在地理空间上是相邻的。但是，面对覆盖面积更大、更复杂的交通网络时，这种方法固然不适用。因此，目前在关于分段预测路段的旅行时间的研究

中，存在着一个问题：如何找到具有相似空间特征的路段，并且针对这些路段进行分段建模来预测它们的旅行时间。

## 1.3 研究内容

通过回顾以上的相关研究，本文发现了目前在分段预测旅行时间的研究领域中仍存在以下问题值得探讨：

(1) 如何有效挖掘路段之间的空间关联信息；

(2) 在空间关联紧密的路段组成的群中，如何预测各路段的旅行时间；

为了解决这些问题，本文提出了基于空间特征层次聚类算法(Spatial Feature-based Hierarchical Clustering, SFHC)和深度多输入门控循环单元网络模型（Deep Multi-input Gated Recurrent Unit, DM-GRU）的旅行时间预测方法。在本方法中，SFHC 算法用于自动挖掘交通网络中具有相似的空间特征、紧密的空间关联性的路段，并把这些路段放在同一个群中。在使用 SFHC 将交通网络分割成多个路段群后，为了更好地利用交通数据的时间、空间依赖和外部信息，本方法采用 DM-GRU 模型预测各路段的旅行时间。经过和五类基准模型、结合了三种经典聚类算法的对照模型比较，本文提出的方法取得了最优的旅行时间预测效果。

本文之后的结构如下：第二章介绍基于 SFHC 和 DM-GRU 的旅行时间预测方法的原理。第三章围绕实验内容和结果展开，而结论和未来的研究方向将在最后一章中呈现。

# 第 2 章 研究方法

本章内容首先介绍基于 SFHC 和 DM-GRU 的旅行时间预测方法的总体框架，其次展现 SFHC 算法的原理，最后将围绕 DM-GRU 结构进行描述。

## 2.1 基于 SFHC 和 DM-GRU 的旅行时间预测方法

基于 SFHC 和 DM-GRU 的旅行时间预测方法的总体框架为：首先使用 SFHC 算法对目标交通网络进行分割，从而得到多个路段群。之后，在每个路段群中建立对应的 DM-GRU 模型用于预测群中每个路段在未来时刻的旅行时间。该框架的示意图如图 2-1 所示。

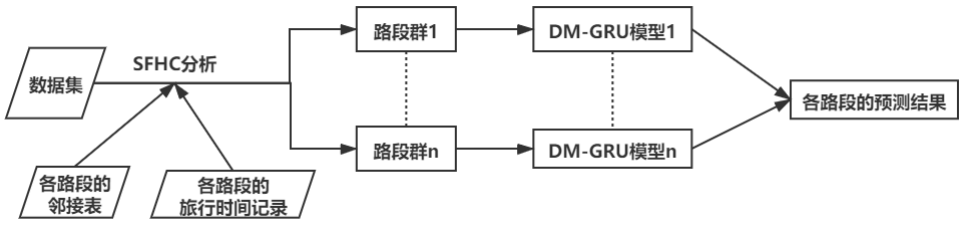



图 2-1. 结合 SFHC 的 DM-GRU 模型

## 2.2 空间特征层次聚类算法

本研究需要找到一个有效的方法来分析路段之间的空间特征相关程度或空间关联性。如果两个路段拥有相似的交通模式以及互相接壤的地理位置，那么它们就被定义为具有相似的空间特征或紧密的空间关联性。

在数据挖掘领域，聚类算法常被用于寻找数据样本之间的关联性。常见的聚类算法有四种类型：划分法、层次法、密度算法和网格法。基于划分法的代表聚类算法为 K 均值算法。该算法使用样本之间的距离来判断它们的相似度，而且需要事先指定聚类分析后得到的群数。而密度算法和网格法都是根据数据空间中某区域内的样本数量进行聚类。层次法采取自底向上的方法，逐一合并当前具有最高相似度的两个候选群，直到已获得规定的群数或者没有任何一对候选集的相似度达到阈值。

为了确定路段之间的空间关联性，本研究首先需要定量分析不同路段在交通状况变化模式方面的相关程度。由于样本之间的距离会受到样本值自身量纲的影响，因此基于距离的相似度度量方法往往无法直观地反映出样本之间的关

联性，所以本研究不采用 K 均值算法进行聚类分析。同时，由于交通数据具有一定的特殊性，所以在密度算法和网格法中难以找到合适的方法用于划分数据空间。因此，本研究最后选择层次法来分析路段之间交通模式的关联程度。

对于旅行时间的预测问题，传统的层次法虽然可以实现定量分析路段之间交通状况的相关程度，但是无法保证群中各路段是相邻的，从而无法获得完整的空间关联性或空间特征。因此，本研究采用空间特征层次聚类算法（Spatial Feature-based Hierarchical Clustering, SFHC）对交通网络进行聚类分析。与传统层次法不同的是，SFHC 的候选群只从相邻的路段中产生。在 SFHC 中，每两个相邻路段被选出并计算得到二者所属群的相似程度后，如果它们的相似度大于阈值就被合并。因此，当遍历整个交通网络之后，SFHC 就可以得到具有相似交通模式并且在地理空间上紧密联系的路段以及这些路段组成的群。

SFCH 算法的目标是生成一个群标记数组，其中每个元素对应一个路段所属的群。实现该算法的具体流程在算法 2-1 中显示。而通过遍历群标记数组，并将具有相同标记的路段放在一个集合中，就可以获得拥有相似旅行时间变化趋势和紧密空间联系的路段群。

SFHC 采用如公式（2-1）所示的平均链接度量方法计算群与群相似度。公式中的$|A|$、$|B|$分别代表群$A$和群$B$包含元素的数目，而$A_i$、$B_j$则是群中的元素。

$$sim(A,B)=\frac{\sum_i^{|A|}\sum_j^{|B|}sim(A_i,B_j)}{|A||B|} \qquad (2-1)$$

而计算元素之间相似度的方法是基于皮尔逊相关系数法。皮尔逊相关系数法的具体计算过程如公式（2-2）所示。对于变量$X$，$\mu_X$、$\sigma_X$分别代表$X$的均值和标准差。

$$\rho(X,Y)=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \qquad (2-2)$$

在许多研究中，皮尔逊相关系数法用于衡量两个对象之间的线性相关程度。如果计算得到的系数大于 0，说明两个变量之间是正相关的关系；而系数为负数时，二者则呈现负相关的关系。而系数的绝对值越接近 1，则说明二者的相关性越强。

**算法 2-1 SFHC**

输入：路段列表$L$,各路段的旅行时间记录$R$, 网络对应的邻接表 $M$, 阈值$T$

输出：群标记数组

输入：路段列表$L$,各路段的旅行时间记录$R$, 网络对应的邻接表 $M$, 阈值$T$

步骤 1：生成每个路段的初始群标记；

步骤 2：获得群标记数组

While $L \neq \emptyset$

{ 从$L$中取出一个路段，记为 $i$;

从$M$中取出$i$的邻居路段列表$N_i$;

While $N_i \neq \emptyset$

{ 从$N_i$中取出一个路段，记为$j$

从群标记数组中取得$i$和$j$各自对应的群：$C_i$ 和 $C_j$

If $C_i$ 和 $C_j$ 不是一个群

{ 计算$C_i$ 和 $C_j$ 的相似度$sim$;

If $sim > T$:

{

把$C_i$ 和 $C_j$ 合并为$C_k$，更新$i$和$j$对应的群为$C_k$

}

令$j = j + 1$(准备取出下一个邻接路段)

}

}

令$i = i + 1$(准备取出下一个路段，检查和其邻接节点的相似度)

}

## 2.3 深度多输入门控循环单元模型

除了挖掘路段之间的空间关联性，本研究还需要找到一种有效利用数据的时间、空间依赖的模型来进行回归分析。

在序列学习中，拥有操控输入开关和遗忘开关的长短期记忆单元（LSTM）[28]被用于解决传统循环神经网络在输入长序列数据时的梯度消失问题。但是由于 LSTM 中的参数较多，所以它需要较长的时间来更新参数、收敛模型。因此，研究[29]提出了一种被认为是 LSTM 简化版本的神经元结构——门控循环单元(Gated Recurrent Unit, GRU)。图 2-2 展示了 GRU 的内部结构，其中三种颜色的箭头分别对应 GRU 中的三个参数$W_r$、$W_z$、$W_h$。图中用于控制重置门$r_t$和更新门$z_t$的信号分别由公式（2-3）和公式（2-4）产生。根据公式（2-5），重置门$r_t$决定了来自上一时间点的信息$h_{t-1}$是否被保留在新的隐藏状态$h_t'$中；而根据公式（2-6），更新门$z_t$被用于同时控制有多少$h_{t-1}$和多少当前的输入信息$x_t$可以被

保存在最终的隐藏状态$h_t$以及输出$y_t$中。$z_t$的信号越大，$h_{t-1}$可以在$h_t$和$y_t$被保存得越多，而$x_t$则被使用得越少。在公式（2-3）至（2-6）中，$\sigma$代表 sigmoid 激活函数；⊙则是一种逐元素的矩阵乘法。

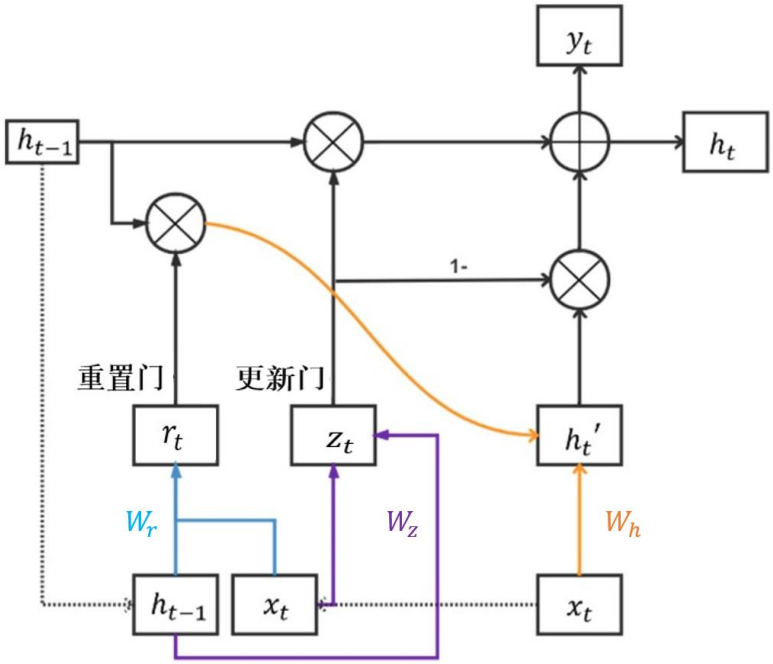



图 2-2. GRU 的结构

$$r_t = \sigma(W_r[h_{t-1}, x_t]^T) \quad (2-3)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t]^T) \quad (2-4)$$

$$h'_t = tanh(W_h[h_{t-1} \odot r_t, x_t]^T) \quad (2-5)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t' \quad (2-6)$$

由以上描述可以看出，GRU 中的门控机制使得历史时间点的信息被灵活保存下来。如果重置门越活跃，那么历史时间点的信息就被遗忘得更多，从而学习到序列的短期依赖；假设更新门越活跃，那么历史时间点的信息就被保留得更多，从而获得序列的长期依赖。在旅行时间的预测中，未来时间点的信息往往与前几个时间点的信息有关，因此使用 GRU 可以选择性地保留旅行时间序列的长期或短期时间依赖。

而在本研究提出的多输入 GRU 结构（Multi-input GRU, M-GRU）中，输入向量由多个路段在同一时刻的旅行时间组成。如图 2-3 所示，GRU 的输入向量$[x_{t,1}, x_{t,2}, \dots, x_{t,n}]$对应$n$个路段在时间点$t$的旅行时间，而输出向量$[x_{t+m,1}, x_{t+m,2}, \dots, x_{t+m,n}]$则对应这些路段经过$m$个时间间隔后的旅行时间。因此，

M-GRU 可以学习多个路段之间的相互影响，从而有效利用了各路段之间的空间依赖。

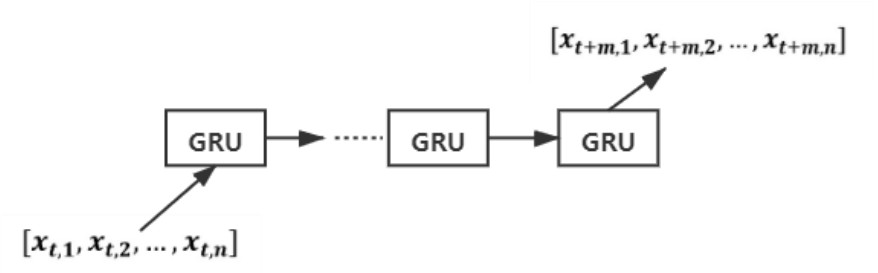



图 2-3. M-GRU

除了时间、空间维度上的特征，获取目标时刻是否处于周末、是否处于交通高峰期等外部信息也有利于预测该时刻的旅行时间。由于目标预测时刻所处的日期和时间段是离散型变量，所以需要先使用独热编码将它们转为类别型变量，然后把它们输入模型。比如，在一周的时间范围内，星期三的独热编码为（0，0，1，0，0，0，0）。但是，这种编码方式造成数据的维度过高、信息量过于稀疏，同时无法提取不同日期或不同时刻之间的联系。例如，一些道路在周六、周日这两天的交通状况是类似的。为了解决这一问题，Liu Yang 等人[30]通过嵌入层技术获取这些外部信息，并将其用于地铁客流量的预测中。嵌入技术是一种在自然语言处理中被广泛用于将高维数据压缩成致密的低维向量的方法，该压缩过程如公式（2-7）所示。对于拥有 n 种类别的变量$x$, $X$为$x$经过独热编码后产生的输入向量，其大小为$1 \times n$。$X$经过大小为$n \times m$的矩阵$W$的线性转换后，可得到$1 \times m$的输出向量$Y$。在实际运用中，$n$一般远大于$m$。

$$Y = X \cdot W \tag{2-7}$$

参照他们的研究成果，本文提出的深度多输入 GRU 模型（Deep Multi-input GRU, DM-GRU）同样使用嵌入层来获得外部信息。DM-GRU 的结构如图 2-4 所示。在时刻$t$，M-GRU 层从输入的时间序列中获取时间依赖信息，得到向量$M_t$; 目标预测的时间点处于一天 24 小时内的具体时间戳作为一种外部信息，先经过独热编码变成高维的类别型变量，再被嵌入层压缩成低维向量$E_t$。最后，$M_t$和$E_t$经过图 2-5 所示的混合过程得到模型的输出向量$[y_{t,1}, y_{t,2}, \cdots, y_{t,i}, \cdots]$，其中$y_{t,i}$代表路段$i$在目标时间点$t$的旅行时间。

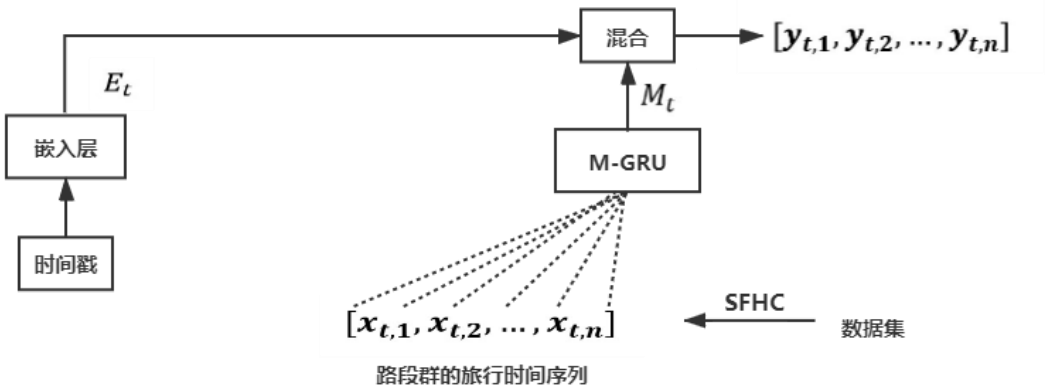



图 2-4. DM-GRU

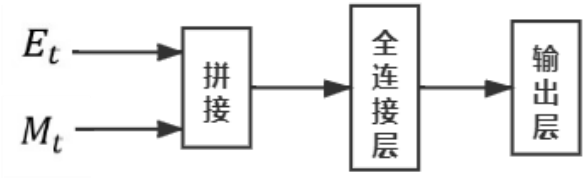



图 2-5. 混合操作

# 第 3 章 实验结果与讨论

本章首先介绍实验数据集的构成以及预处理数据的方法。随后的实验设置部分将介绍各模型和聚类算法的参数设定。关于分别结合了不同聚类算法的基准模型和 DM-GRU 模型的实验结果在第 4 节中呈现。而本章的结尾部分围绕这些实验结果进行相关的讨论。

## 3.1 数据说明

本研究基于福州大学土木学院提供的福建漳州市区旅行时间数据集进行相关的实验。该数据集覆盖了 300 个路段从 2018 年 4 月 30 日到 2018 年 6 月 29 日，为期 60 天的真实旅行时间。这些时间记录的间隔大约为一分钟。经过统计各路段的数据分布情况，大多数路段在晚上 11 点至次日上午 7 点的时段中的旅行时间记录较少，同时仅有 82 个路段拥有低于 30%的缺失记录。因此，后续实验将围绕这 82 个路段在 41 个工作日（除去公共假期）中从 8 点至 23 点的数据。同时，本实验从这些路段中挑选了较拥塞的路段以及它们周围的路段。这里需要说明的是，如果一条路段时速小等于每小时 20 公里的记录超过 50%，则定义该路段为拥堵的。通过以上挑选路段的算法，本实验最后筛选得到 16 个路段，从而得到一个较拥堵的交通网络。该交通网络在地图中的位置如图 3-1 所示。其中每一条带有箭头的线段表示一个路段，而箭头的指向为路段的朝向，而红色线段代表拥堵路段。


图 3-1. 实验交通区域

根据常规做法，本实验按照 6:2:2 的比例将数据集划分成训练集、验证集、测试集。还需要注意的是，对于时间序列的回归问题，为了避免数据泄露，本实验按照时间顺序划分得到这三个数据集[31]。

## 3.2 数据预处理

根据旅行时间预测研究领域的常规做法，实验数据需要被整合成间隔为 5 分钟的新集合。具体的整合过程为：

步骤 1. 在同一天内，使用前一个时间点的值来填补当前的缺失值。如果在 8:00 出现缺失值，将根据该路段的限速来决定其最短旅行时间，并使用这个时间来填补 8:00 时间点的值。

步骤 2. 由于原始数据的间隔为 1 分钟，所以对每 5 笔数据取均值来代表路段在该 5 分钟内的平均旅行时间。

经由以上过程，每个路段在新集合中拥有 60×180 笔旅行时间记录,其中 60、180 分别为数据集覆盖的天数和每天的旅行时间记录数。

此外，研究[32][33]表明，对数据进行平滑处理，可以有效减少数据中的噪声或者异常值对于后续聚类分析的干扰。因此，本实验在新集合上进行指数移动平均法(Exponential Moving Average, EMA)的平滑操作。公式（3-1）代表 EMA 的计算过程。$EMA_t$、$EMA_{t-1}$分别代表当前时间点以及上一时间点被平滑后的值，$Value_t$代表当前时间点的数值。

$$EMA_t = \alpha \times Value_t + (1-\alpha) \times EMA_{t-1} \qquad (3-1)$$

而$\alpha$是移动平滑系数，其计算过程如公式（3-2）所示。其中$n$为用于平滑数据的时期长度。在本实验中，$n$设置为 6，即以前 30 分钟的历史数据来平滑当下时间点的值。

$$\alpha = \frac{2}{n+1} \qquad (3-2)$$

某路段一天内的旅行时间经过 EMA 平滑后的效果如图 3-2 所示。图中绿色曲线对应原始的时间序列，而红色线条代表经过 EMA 平滑后的数据。可以看出，平滑操作显著地减少了数据的波动程度。

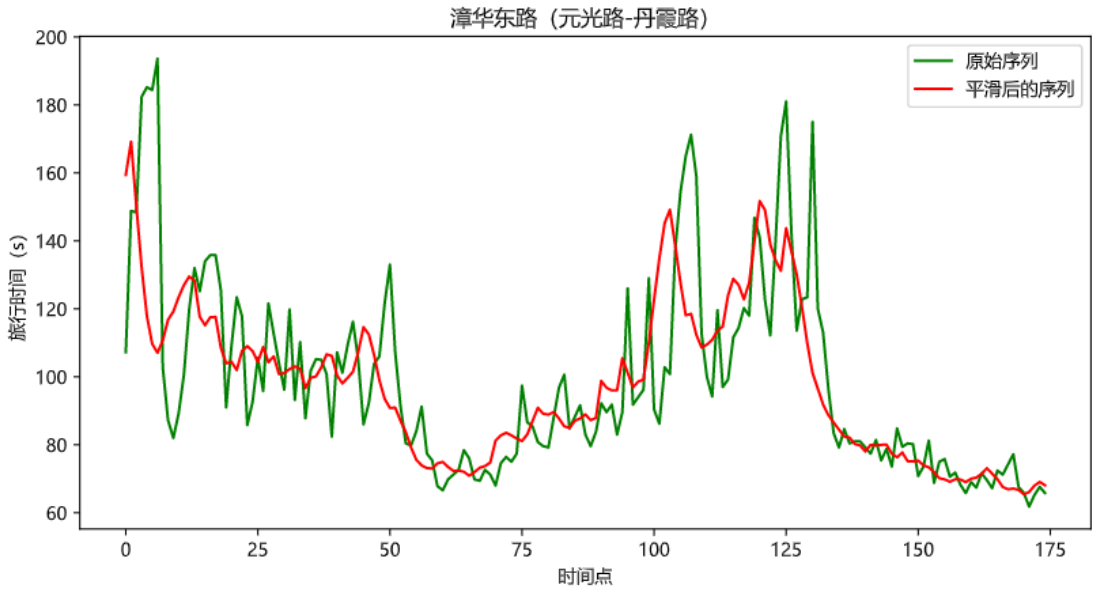



图 3-2. EMA 平滑效果

最后，为了加快模型在训练时的收敛速度，需要对数据集进行归一化。本实验采取最大-最小值归一化的方法，该方法原理如公式（3-3）所示。式中的$x$为原始值，$x^*$代表归一化后的值。由于各路段的长度不同，所以需要对每个路段的数据分别进行归一化，并且各路段的最大、最小旅行时间$x_{max}$、$x_{min}$皆来源于训练集数据。在测试集中，同样使用这些$x_{max}$、$x_{min}$并根据公式（3-3）对数据进行归一化。

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (3-3)$$

## 3.3 实验设置

### 3.3.1 模型和算法参数

#### 3.3.1.1 聚类算法

经过尝试 4 种不同的阈值，并根据结合 SFHC 的 DM-GRU 模型的测试结果，得到了本实验中 SFHC 的最佳阈值 0.7，即两个候选群的旅行时间变化趋势的相似度需达到 70%以上，才可以被合并。寻找最佳阈值的实验结果如表 3-2 所示，其中 SFHC 阈值设置为 0.7 时， 结合了 SFHC 的 DM-GRU 模型取得了最低的预测误差。

此外，为了探究 SFHC 相对于其他聚类算法，是否可以有效降低旅行时间预测误差，本实验设置了采用原始层次聚类、K 均值聚类和谱聚类算法的对照

实验。谱聚类是一种基于图论的算法，它将样本看作无向权重的图，并通过切图的方式，使得子图之间的权重和尽量小，而子图内部的点之间的权重和尽可能大，从而划分出不同的群[34]。在本实验中，图中各边的权重为不同路段之间的相似度。因此，使用该聚类算法也可以实现定量分析路段之间的相关程度。

在设置对照组实验时，先使用 SFHC 进行聚类分析，然后将原始层次聚类、K 均值和谱聚类的目标群数设置为 SFHC 得到的群数。此外，还有一组实验将原始层次聚类算法的阈值设置为 0.7，和 SFHC 一致。

表 3-2 结合不同阈值的 SFHC 的 DM-GRU 的测试结果

| SFHC 阈值 | 0.6 | 0.65 | 0.7 | 1 |
|---|---|---|---|---|
| DM-GRU | 3.68% | 3.43% | **3.42%** | 3.43% |

3.3.1.2 DM-GRU

在 DM-GRU 中，嵌入层的输出单元为 3， M-GRU 的输出向量为 16 维，输出层的输出单元数和目标路段群中的路段数相同。输出层的激活函数是线性的。

3.3.1.3 基准模型

本实验的基准模型选取了支持回归向量（SVR）、线性回归模型（LR）、多层感知器（MLP）、基于时空矩阵的卷积神经网络（CNN）[11]以及使用注意力机制捕捉时间序列的周期性和迁移特征的时空动态网络模型（STDN）[9]。由于本文的研究对象以及实验数据的结构和研究[9][11]中的内容并非完全相同，所以本实验构建的 CNN 和 STDN 和它们在原论文中的结构有区别，但是保留了这两种最新模型的核心部分。这些基准模型的隐藏层、输出层的激活函数和 DM-GRU 相同。

根据以上五类基准模型在验证集中的表现，将它们的参数设置如下：MLP 采用两个分别包含 32、24 个输出单元的隐藏层；CNN 中包含了设置一个 3×3 大小的卷积核；而本实验的 STDN 模型同样采取 M-GRU 结构，每个 M-GRU 的输出单元设置为 8，并且根据原论文的实验设置，本实验选取当前时间点在前 3 天同期时刻前后 30 分钟的数据作为长期时间依赖信息。

### 3.3.2 评价指标

在评价模型表现阶段，需要先对模型输出进行反归一化，从而得到真实量纲下的旅行时间。但是由于越长的路段往往具备更长的旅行时间，同时也可能

伴随着更大的旅行时间预测误差。如果直接使用反归一化后的值来评估模型的表现，会导致同一模型对于不同长度的路段的旅行时间预测误差的范围有所差异，从而影响模型整体表现的评估。比如在某交通网络中，长路段的旅行时间预测误差范围在 5 至 10 秒内，而短路段的误差范围在 2 到 3 秒内，那么该网络的平均误差可能会大于 3 秒，从而无法反映出模型对于短路段的真实预测效果。因此，为了更好地查看模型的表现，实验的评估阶段需要消除路段的长度对于评价结果带来的影响。因此，本实验选取平均绝对百分比误差（MAPE）作为模型的评价指标。MAPE 的计算过程如公式（3-4）所示。其中$y$为真实的旅行时间,$\hat{y}$指的是由模型推理得到的值，$N$为样本数量。该评价指标可以衡量实际值和预测值之间的差异相对于实际值的大小。

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{\hat{y} - y}{y} \right|}{N} \times 100 \qquad (3-4)$$

### 3.3.3 实验环境

本实验使用 Python 3.6.8 中基于 Tensorflow 后端的 Keras 库和 sklearn 库进行模型的相关操作，同时使用 Numpy 库进行数据处理。训练模型的时候，本实验借助英伟达 Tesla P100 的 GPU 进行运算。

## 3.4 实验步骤

本实验的过程由以下三个步骤组成：

步骤 1. 聚类分析

步骤 2. 构造模型的输入、输出向量

步骤 3：训练、测试模型

由于在基准模型和 DM-GRU 模型的实验过程中，除了模型不同，其余步骤都是类似的，因此本节在介绍每个步骤的内容时，忽略了具体模型的名称。

### 3.4.1 聚类分析

SFHC 和原始层次聚类的输入向量为$[\overline{x_{i,1}}, \overline{x_{i,2}}, \dots, \overline{x_{i,j}}, \dots]$，其中每个元素$\overline{x_{i,j}}$代表路段$i$在一天中的时间点$j$的旅行时间，而该值是路段$i$在训练集 48 天中对应时间点$j$的平均旅行时间，其计算过程如公式（3-5）所示。式中的$x_{i,d,j}$代表路段$i$在训练集中第$d$天的时间点$j$的旅行时间。使用这种方式得到的向量可以反映

出一个路段在训练集中整体上的旅行时间变化情况。谱聚类算法的输入是基于前面两种算法的输入而构建的一个相似度矩阵。该矩阵的结构如公式（3-6）所示，其中每个元素$sim(i,j)$代表路段$i$和路段$j$之间的相似度。对角线上的1代表各路段和其本身的旅行时间变化情况重合。同时，由于该矩阵是对称的，所以式中用省略号代表对称的元素。需要注意的是，由于谱聚类的输入要求为非负矩阵，所以本实验将值为负数的相似度调整为 0。

得到以上输入后，就可以使用对应的算法进行聚类分析。

$$\overline{x_{i,J}} = \frac{\sum_{d=1}^{48} x_{i,d,j}}{48} \tag{3-5}$$

$$M = \begin{bmatrix} 1 & sim(1,2) & sim(1,3)\cdots & sim(1,n) \\ \vdots & 1 & sim(i,j)\cdots & sim(i,n) \\ \vdots & \ddots & 1\cdots & \vdots \\ \vdots & \cdots & \cdots & 1 \end{bmatrix} \tag{3-6}$$

### 3.4.2 构造模型的输入、输出向量

SVR、MLP 和 LR 的输入向量由公式（3-7）所示。向量中每个元素$x_{i,j}$代表路段$j$在时刻$i$的旅行时间。因此，这两种模型的输入是由各路段在目标预测时刻的前$m$个时间点的旅行时间组成。CNN、STDN 和 DM-GRU 的输入矩阵如公式（3-8）所示。其中$x_{i,j}$同样代表路段$i$在时刻$j$的旅行时间。矩阵中的行向量代表某时刻所有路段的旅行时间集合，如$[x_{t-m,1}, x_{t-m,2}, \dots, x_{t-m,n}]$表示在目标预测时刻$t$之前的第$m$个时间点$n$个路段的旅行时间序列。

虽然这些模型的输入结构不同，但是它们的输出都为$n$维的向量，如公式（3-9）所示。其中每个元素代表对应路段在目标预测时刻的旅行时间，如$y_n$表示路段$n$的旅行时间。

$$X = [x_{t-m,1}, x_{t-m,1}, \dots, x_{t-1,1}, \dots, x_{t-m,n}, x_{i,j}, \dots, x_{t-1,n}] \tag{3-7}$$

$$X = \begin{bmatrix} x_{t-m,1}, x_{t-m,2}, \dots, x_{i,j}, \dots, x_{t-m,n} \\ x_{t-(m-1),1}, x_{t-(m-1),2}, \dots, x_{i,j}, \dots, x_{t-(m-1),n} \\ \cdot \\ \cdot \\ \cdot \\ x_{t-1,1}, x_{t-1,2}, \dots, x_{i,j}, \dots, x_{t-1,n} \end{bmatrix} \tag{3-8}$$

$$Y = [y_1, y_2, \dots, y_n] \tag{3-9}$$

本实验的目标是：利用前 20 分钟的历史旅行时间来预测未来 5 分钟的旅行

时间，即以当前时间点前4个时间点的旅行时间来估计未来1个时间点的值（本实验中相邻时间点的间隔为 5 分钟）。因此，在此步骤中，根据不同模型的结构和以上对应的公式，将预处理后的数据集进行维度变化。变化后，各模型的输入输出向量的维度如表 3-1 所示。表中的$n$为目标路段群所包含路段的数量。

表 3-1. 各模型输入和输出向量的维度

| | 输入 | 输出 |
|---|---|---|
| **SVR、MLP、LR** | $n \times 4$ | $n$ |
| **CNN、STDN、DM-GRU** | $(4, n)$ | $n$ |

### 3.4.3 训练、测试模型

该步骤的内容包含训练和测试模型两部分，如图 3-3 所示。

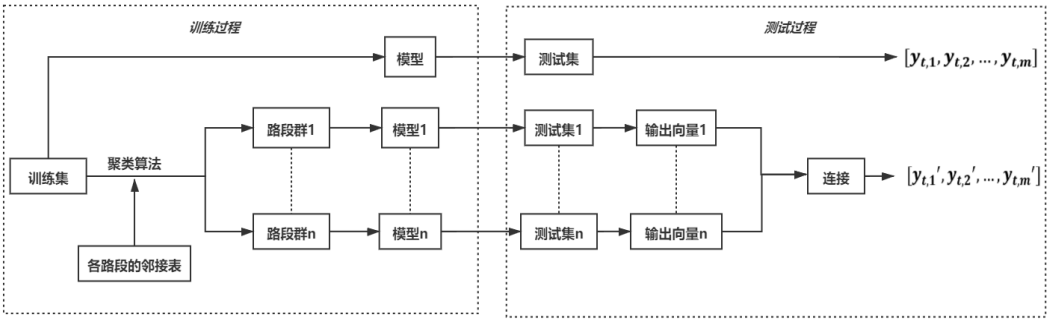



图 3-3. 训练、测试模型的过程

#### 3.4.3.1 模型的训练

在训练模型的过程中有以下步骤：

① 在总的训练集中训练模型；

② 使用聚类算法将交通网络划分为多个路段群。由于每个路段群中路段数量的不同，需要按照步骤①重新构造输入、输出向量。之后，为每个路段群建立结构相同的回归模型，并进行训练。

在训练过程中，模型的损失函数为最小平方误差（MSE），其计算过程如公式（3-9）所示。此外，本实验选择 Adam 作为优化器，并设置训练迭代回合为 100 轮。同时，为了防止模型出现过拟合的状况，在训练模型的过程中加入了回调函数 early stopping 来监听验证集上损失函数值的变化。当验证集上的损失函数值不再减小的次数达到 10 次后，模型即停止训练。

$$MSE = \frac{\sum_{i=1}^{N}(\hat{y} - y)^2}{N} \quad (3-9)$$

#### 3.4.3.2 模型的测试

在测试集上运行所有模型，并获得各自的测试结果。在图（3-3）中，总数据集对应的模型的输出为$[y_1, y_2, \cdots, y_m]$, 为路段数量。而获得每个路段群对应模型的输出向量之后，将它们连接在一起，从而得到所有路段在分段建模后的测试结果$[y_1', y_2', , \cdots, y_m']$。

## 3.5 实验结果

### 3.5.1 聚类分析结果

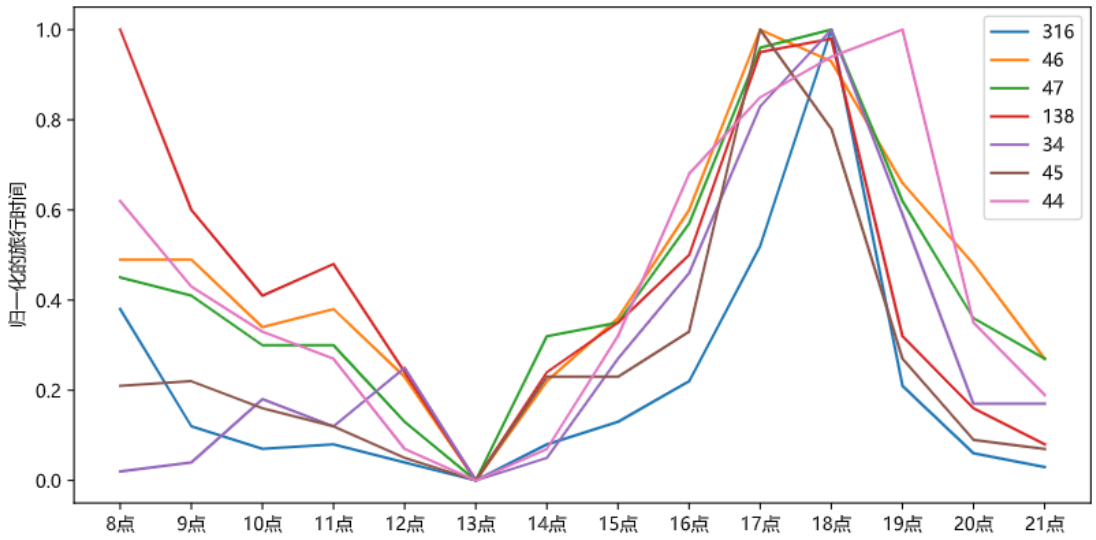
#### 3.5.1.1 SFHC

将 SFHC 的聚类结果导入地图后得到图 3-4 所示的效果，图中不同的颜色代表不同的路段群。可以看出，每个路段群中的各路段是相互接壤的。

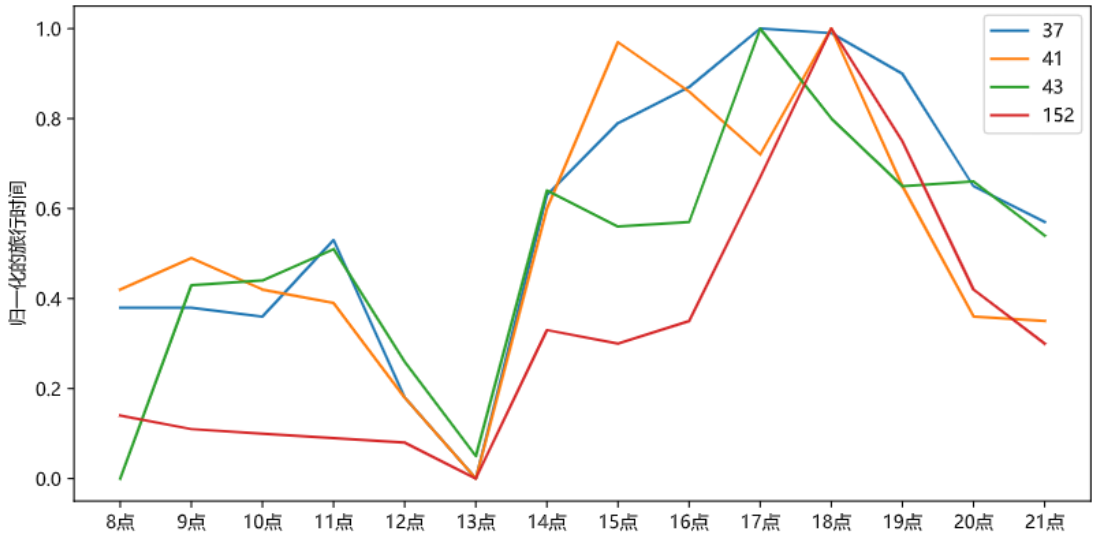

图 3-4. SFHC 聚类结果可视化

接下来的内容将围绕两个路段群来探讨 SFHC 算法是否可以提取具有相似交通模式的不同路段，并且探究导致二者的交通模式不同的原因。这里需要注意的是，为了更清楚地显示各路段在一天中的旅行时间变化趋势，本部分在可视化聚类分析的结果时，对时间记录进行了最大-最小值归一化。图 3-5 中每个子图的纵轴代表归一化后的旅行时间，横轴是一天中的时间点，而不同颜色的线条对应不同路段的通行状况变化。如果曲线在某时刻对应纵轴的值接近 1，

则说明此时的旅行时间靠近一天中的最高点。反之，纵轴对应的值接近 0，说明对应的旅行时间靠近最低点。




$(a)$




$(b)$

图 3-5. SFHC 分割得到的路段群对应的旅行时间变化趋势

1) 图 3-5($a$)对应的路段群由图 3-6 中的红色线条标明。由图中可以看出，这些路段集中在生活小区、学校和医院附近。这些区域的通行高峰一般集中于上午上学和上班的时间段以及下午放学和下班的时段。而从图 3-5($a$)的旅行时间变化曲线可以看出，这些路段在上午 8-9 点和下午 5-7 点有明显的旅行时间高峰。
2) 图 3-5($b$) 对应的路段群由图 3-7 中的紫色线条标明。由图中可以看出，这些

路段集中在饮食场所、公园等娱乐休闲场所附近。这些区域的交通高峰时期一般集中于大众下班或是晚上的时候。所以，在图 3-5($b$)中可以看到这些路段上午时段的交通状况比较通畅，而在下午 5 点之后，它们的旅行时间基本维持在较高值。

从以上的案例分析可以看出，SFHC 可以找到因处于相同城市功能区域而产生相似交通模式的相邻路段。

#### 3.5.1.2 其他聚类算法

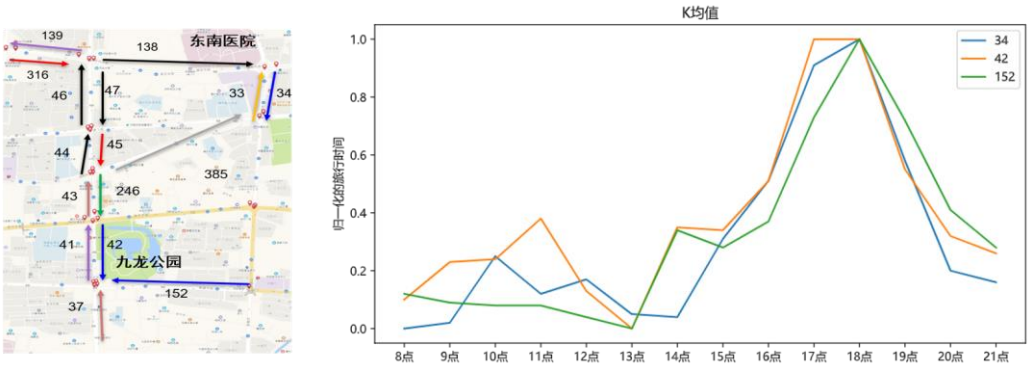
其余聚类算法的分析结果在图3-8中显示。在每个子图中，左图代表导入地图的结果，其中不同颜色代表不同的路段群；而右图则是从对应聚类结果中任意挑选的一个路段群的旅行时间变化趋势图。图 3-8($a$)到图 3-8($d$) 分别为 K 均值、谱聚类、设定目标群数的层次聚类算法（HC-群数）以及设定阈值的层次聚类算法（HC-阈值）的分析结果。

可以看出，在由这些聚类算法得到的路段群中，虽然各路段具有相似的旅行时间变化趋势，但是并非所有路段都是相互接壤的。

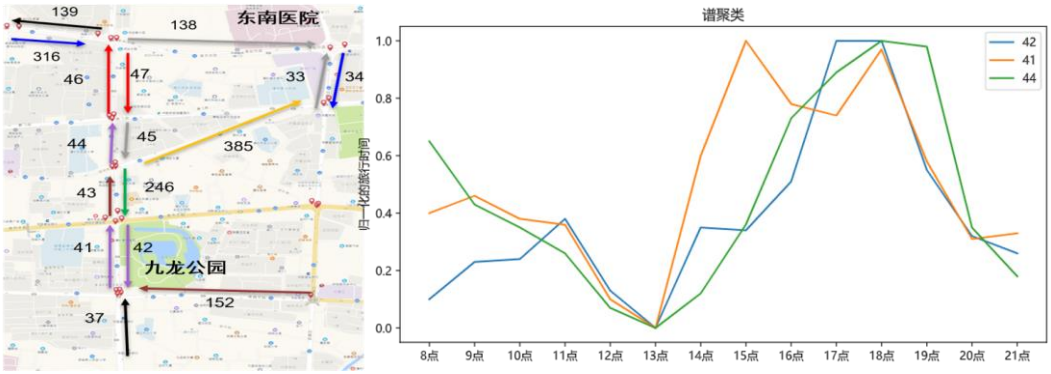

图 3-6


九龙公园
悦华城市广场
世纪广场

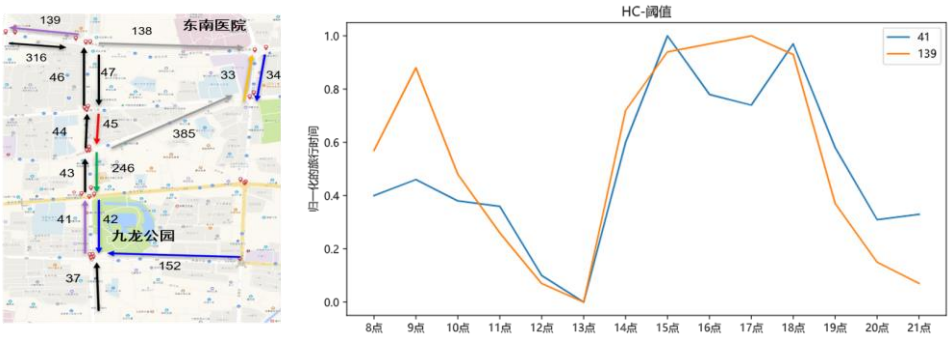
图 3-7


东南医院
九龙公园
K均值
归一化的旅行时间
34
42
152

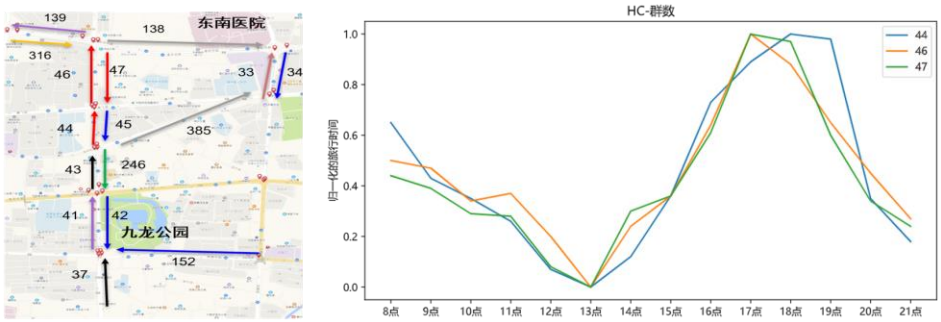
$(a)$


东南医院
九龙公园
谱聚类
归一化的旅行时间
42
41
44

$(b)$




(c)




(d)

图 3-8.其他聚类分析方法的可视化结果

## 3.5.2 模型测试结果

表 3-3 显示了各模型以及它们结合不同聚类算法的测试结果。每个模型在测试集上得到的最小预测误差用粗体标明。而所有实验组中的最小误差用粗体和下划线标明。从表 3-3 可见，基于 SFHC 的 DM-GRU 模型在预测旅行时间方面的表现（右下角）优于所有基准模型以及这些模型结合其他聚类算法产生的对照模型。

表 3-3 模型测试结果

| | 无聚类 | HC-阈值 | HC-群数 | K 均值 | 谱聚类 | SFHC |
|---|---|---|---|---|---|---|
| **SVR** | 18.11% | 17.95% | 17.45% | **16.59%** | 17.38% | 17.56% |
| **STDN** | 8.69% | 6.02% | 4.79% | 4.74% | **4.48%** | 5.34% |
| **CNN** | 4.45% | 4.04% | 4.14% | 4.16% | 4.54% | **3.99%** |
| **MLP** | 4.5% | 3.76% | 3.99% | 3.98% | **3.67%** | 3.87% |
| **LR** | 3.67% | 3.62% | 3.52% | **3.48%** | 3.51% | 3.54% |
| **DM-GRU** | 3.66% | 3.62% | 3.52% | 3.58% | 3.81% | **3.42%** |

图 3-9 显示了针对某路段，误差最低的两个模型（DM-GRU 和线性回归）结合 SFHC 后的预测效果。红色曲线为该路段在某天的真实旅行时间序列，而不同模型的预测效果用不同颜色的曲线表示。从图中可见，结合 SFHC 的 DM-GRU 模型相对于基准模型可以更好地拟合真实的时间序列变化趋势。

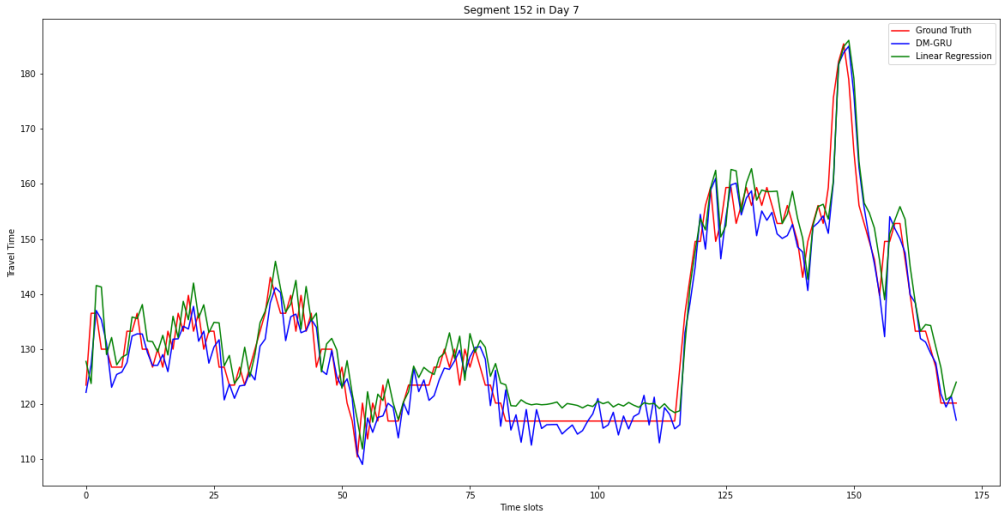



图 3-9 不同模型结合 SFHC 后的预测结果

## 3.6 关于实验结果的讨论

从聚类分析的结果可以看出，SFHC 既可以挖掘出具有处于相同城市功能区、相似交通状况变化趋势的路段，也可以保证了这些路段是相邻的。

接下来的内容围绕模型的测试结果展开讨论：首先，四种聚类算法基本上都可以降低各模型的旅行时间预测误差。这说明了分段建模的方法在旅行时间预测方面的积极作用。其次，在分段建模的过程中，结合 SFHC 的 DM-GRU 模型取得了最低的预测误差，这说明本研究提出的旅行时间预测方法的有效性。此外，对于 SVR、MLP、STDN 和 LR，虽然 SFHC 的作用会差于部分对照算法，但是 SFHC 不要求事先给定目标聚类群数，而是通过规定相似度阈值来直观、定量地控制聚类的程度。与此同时，相比于同样使用阈值的传统层次聚类算法（HC-阈值），SFHC 基本上可以更多地提高各模型的预测准确度。因此，SFHC 仍然具有一定的应用价值。

# 结论

为了充分而便捷地挖掘路段之间的空间关联特征，并且根据挖掘得到的信息进行分段预测旅行时间，本文提出了基于 SFHC 和 DM-GRU 的旅行时间预测方法。在该方法中，SFHC 算法用于寻找交通网络中具有相似旅行时间变化趋势，并且在地理空间上相互连接的路段，从而为分段建立 DM-GRU 模型打下良好的基础。而 DM-GRU 模型在捕捉时间依赖的基础上，再通过同时输入多个路段的旅行时间，可以学习到空间依赖，并且有效利用了外部信息。实验结果证明，相比于五类基准模型和结合了其他三种聚类算法的对照模型，本文提出的方法取得了最少的旅行时间预测误差。因此，该方法对于解决地图导航、智能交通系统中的路线规划问题，有着重要的参考价值。

未来的研究可以尝试在本文提出的方法中加入图神经网络[35][36]来更充分地利用 SFHC 算法提取出的空间信息。此外，在本文提出的方法中，也可以进一步结合基于时间特征的分段建模方式，从而实现同时基于时空特征的分段预测模型。

# 参考文献

[1] XIE P, LI T, LIU J, et al. Urban flow prediction from spatiotemporal data using machine learning: A survey[J]. Information Fusion, 2020, 59.

[2] WILLIAMS B, DURVASULA P, BROWN D. Urban freeway traffic flow prediction - Application of seasonal autoregressive integrated moving average and exponential smoothing models [J]. Transportation Research Record, 1998, 1644:132-141.

[3] SUN S, ZHANG C, YU G. A bayesian network approach to traffic flow forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(1):124-132.

[4] QI Y, ISHAK S. A Hidden Markov Model for short term prediction of traffic conditions on freeways[J]. TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES, 2014, 43(SI):95-111.

[5] WANG J, SHI Q. Short-term traffic speed forecasting hybrid model based on Chaos-Wavelet Analysis-Support Vector Machine theory[J]. Transportation research, Part C. Emerging technologies, 2013.

[6] MA X, TAO Z, Wang Y, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data[J]. Transportation Research Part C: Emerging Technologies, 2015, 54:187-197.

[7] DUAN Y, LV Y, WANG F Y. Travel time prediction with LSTM neural network[C]// IEEE International Conference on Intelligent Transportation Systems. IEEE, 2016.

[8] WANG J, CHEN R, HE Z. Traffic speed prediction for urban transportation network: A path based deep learning approach[J]. Transportation Research Part C: Emerging Technologies, 2019, 100:372-385.

[9] YAO H, TANG X, WEI H, et al. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33:5668-5675.

[10] CHEN M, YU X, LIU Y. PCNN: Deep Convolutional Networks for Short-term Traffic Congestion Prediction[J]. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 2020, 19(11): 3550-3559.

[11] MA X, ZHUANG D, HE Z, et al. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction[J]. Sensors, 2017, 17(4):818.

[12] ZHANG J, YU Z, QI D, et al. Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks[J]. Artificial Intelligence, 2017, 259.

[13] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[C]// ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 28 (NIPS 2015), arXiv: 1506.04214.

[14] NC PETERSON, F RODRIGUES, FC PEREIRA. Multi-output bus travel time prediction with convolutional LSTM neural network[J]. Expert systems with applications, 2019, 120: 426-435.

[15] YUAN Z, XUN Z, YANG T. Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data[C]// the 24th ACM SIGKDD International Conference. ACM, 2018.

[16] RYAN S E, PORTH L S, PORTH S. A Tutorial on the Piecewise Regression Approach Applied to Bedload Transport Data[J]. Usda Forest Service General Technical Report Rmrs Gtr, 2007(189).

[17] MALLICK T, BALAPRAKASH P, RASK E, et al. Graph-Partitioning-Based Diffusion Convolutional Recurrent Neural Network for Large-Scale Traffic Forecasting[J]. Transportation Research Record Journal of the Transportation Research Board, 2020, 2674(2):036119812093001.

[18] STUTZ C, RUNKLER T A. Classification and prediction of road traffic using application-specific fuzzy clustering[J]. Fuzzy Systems IEEE Transactions on, 2002, 10(3):297-308.

[19] ZHENG P, MCDONALD M. Estimation of travel time using fuzzy clustering method[J]. Iet Intelligent Transport Systems, 2009, 3(1):77-86.

[20] 杨世坚, 贺国光. 基于模糊 C 均值聚类和神经网络的短时交通流预测方法[J]. 系统工程, 2004(08):83-86.

[21] ELHENAWY M, CHEN H, RAKHA H A. Dynamic travel time prediction using data clustering and genetic programming[J]. Transportation Research Part C, 2014, 42(may):82-98.

[22] FENG B, XU J, LIN Y, et al. A Period-Specific Combined Traffic Flow Prediction Based on Travel Speed Clustering[J]. IEEE ACCESS, 2020, 8: 85880-85889

[23] LIU D, TANG L, SHEN G, et al. Traffic Speed Prediction: An Attention-Based Method[J]. Sensors, 2019, 19(18):3836.

[24] SHEN G, CHEN C, PAN Q, et al. Research on Traffic Speed Prediction by Temporal Clustering Analysis and Convolutional Neural Network with Deformable Kernels [J]. IEEE Access, 2018, PP:1-1.

[25] CHEN CHI-HUA, HWANG FENG-JANG, et al. Travel Time Prediction System Based on Data Clustering for Waste Collection Vehicles[J]. IEICE Transactions on Information and Systems, 2019, E102.D(7):1374-1383.

[26] PARK D, RILETT L. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks[J]. Transportation Research Record Journal of the Transportation Research Board, 1998, 1617:163-170.

[27] XU M, GUO K, FANG J, et al. Utilizing Artificial Neural Network in GPS-equipped Probe Vehicles Data Based Travel Time Estimation[J]. IEEE Access, 2019, PP(99):1-1.

[28] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[29] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[30] LIU Y, LIU Z, JIA R. DeepPF: A deep learning based architecture for metro passenger flow prediction[J]. Transportation Research Part C: Emerging Technologies, 2019, 101(APR.):18-34.

[31] KAUFMAN S, ROSSET S, PERLICH C. Leakage in data mining: formulation, detection, and avoidance.[J]. ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, 2012, 6(4).

[32] GUARDIOLA I G, LEON T, MALLOR F. A functional approach to monitor and recognize patterns of daily traffic profiles[J]. Transportation Research Part B, 2014, 65(JUL.):119-136.

[33] HITCHCOCK DB, BOOTH JG, CASELLA G. The effect of pre-smoothing functional data on cluster analysis[J]. JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2007,77(11-12): 1089-1101.

[34] LUXBURG U V. A Tutorial on Spectral Clustering[J]. Statistics and Computing, 2004, 17(4):395-416.

[35] LI Y, YU R, SHAHABI C, et al. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting[J]. 2017.

[36] ZHAO L, SONG Y, ZHANG C, et al. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, PP(99):1-11.

# 致谢

在此特地鸣谢福州大学数学与计算机科学学院的陈志华教授对于本研究的全程指导和支持。从研究初期的数据清洗到后期的模型选择的过程中，本人遇到了许多问题，实验进度多次停滞不前。但是，陈教授给予的悉心指导和专业的学术建议，帮助我逐步克服了这些困难，完成本项研究和本论文的撰写。此外，本人也感谢几位研究生学长、学姐的支持。在处理实验数据集的初期阶段，土木学院的许梦云学姐提供了许多信息帮助我熟悉数据集的构成；在复现各论文里的模型阶段，郭灿阳学长对于解决部分问题的方案给予了有用的建议；在实验阶段，方昊和张翊卓两位学长提供了关于使用远程服务器以及 GPU 计算资源的相关建议。