基于信令数据的人口分布预测模型构建

项目提供方：中国联通智慧足迹科技有限公司

第九组：肖飞宇、韦承志、赵嘉欣、张玉生、沈磊、李司棋、荆科

目录

CONTENTS

1 项目背景和需求
Background of the Project

2 模型构建
Establish the Model

3 预测成果分析
Predict and Results

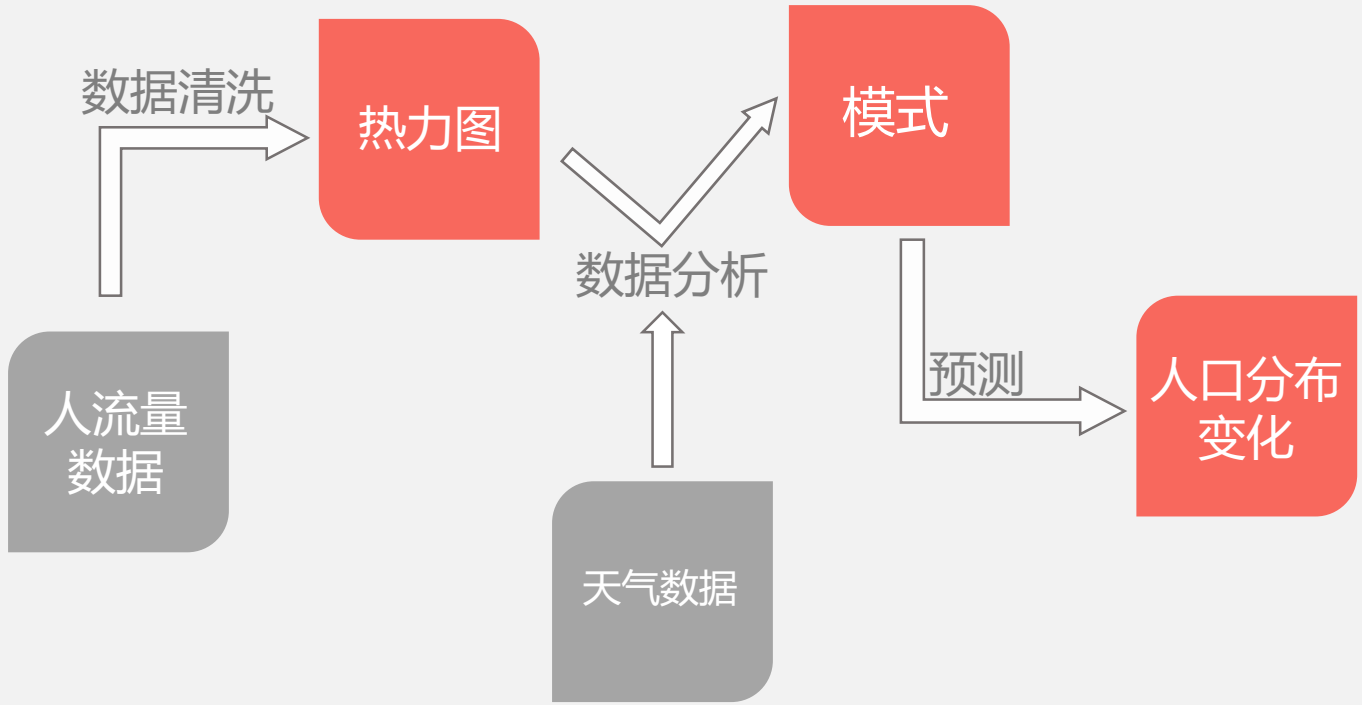
1 项目背景和需求

Background of the Project

项目背景与需求

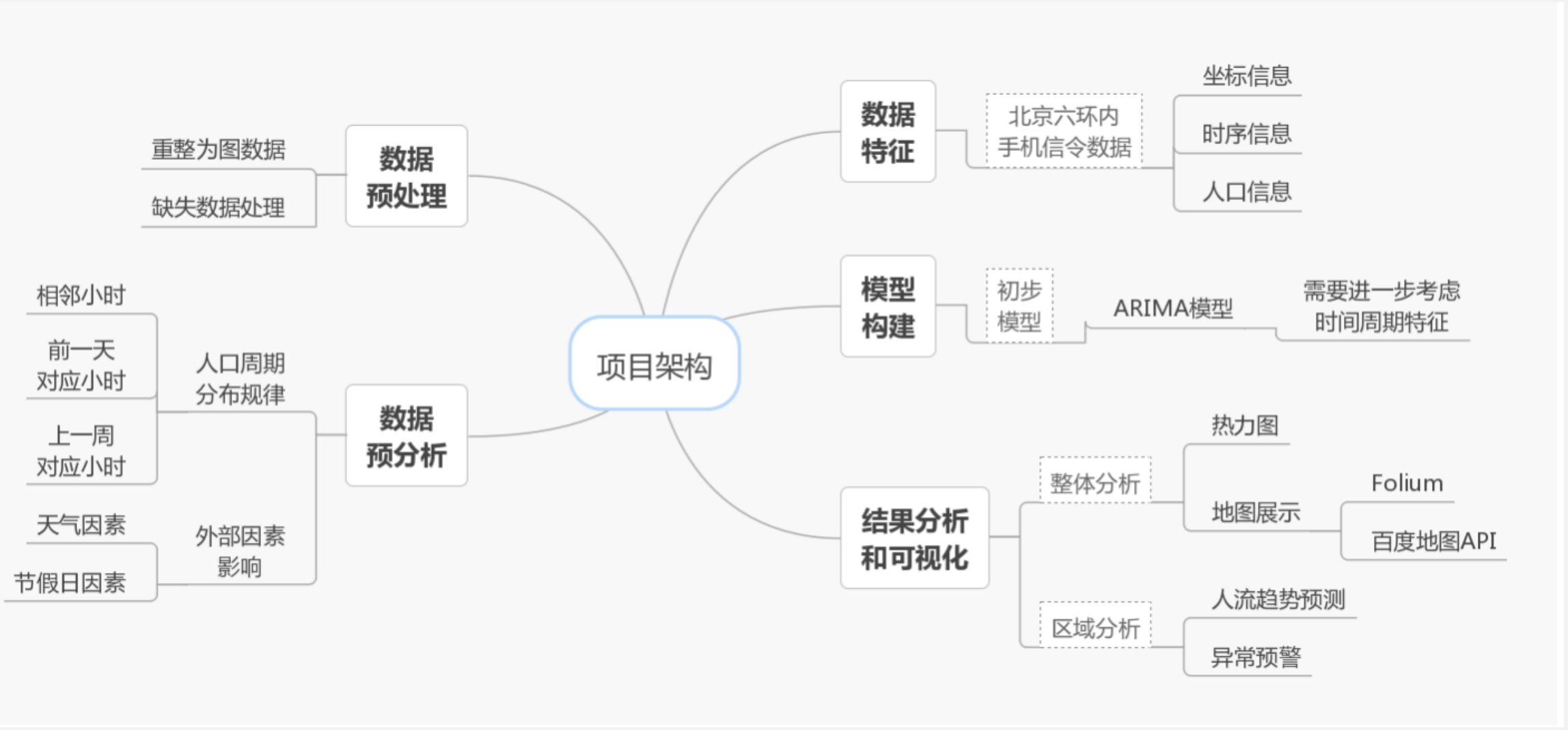
城市区域人流预测

将城市分割成均匀网格，基于交通、气象、时间和事件等多源信息，采用深度学习的方法，来综合预测未来每个网格人口的流动变化从而预测北京六环内人口流动。



技术路线

Routine



2 模型构建

Establish the Model

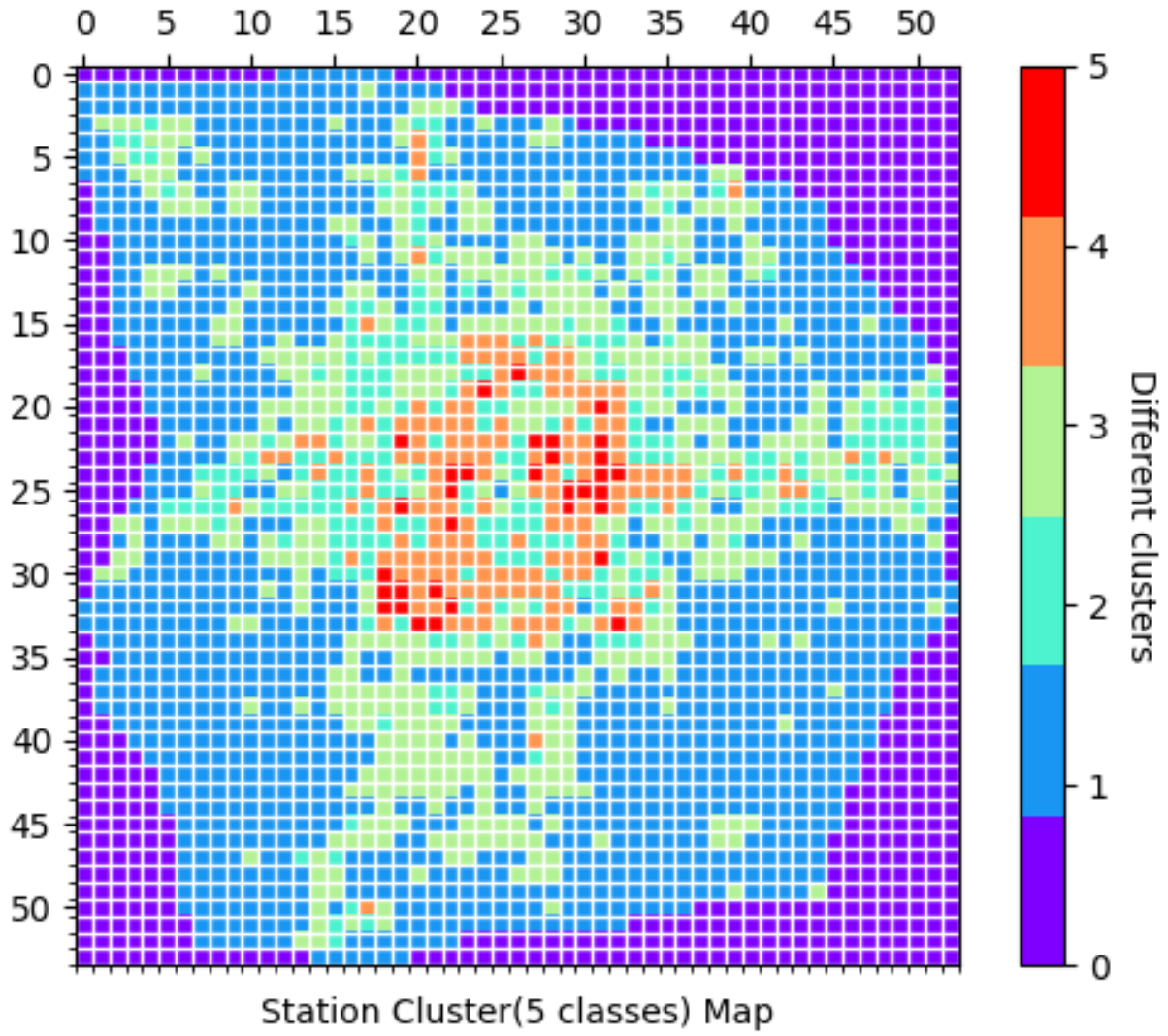
数据预处理

数据结构转变
将手机信令数据转化为图数据

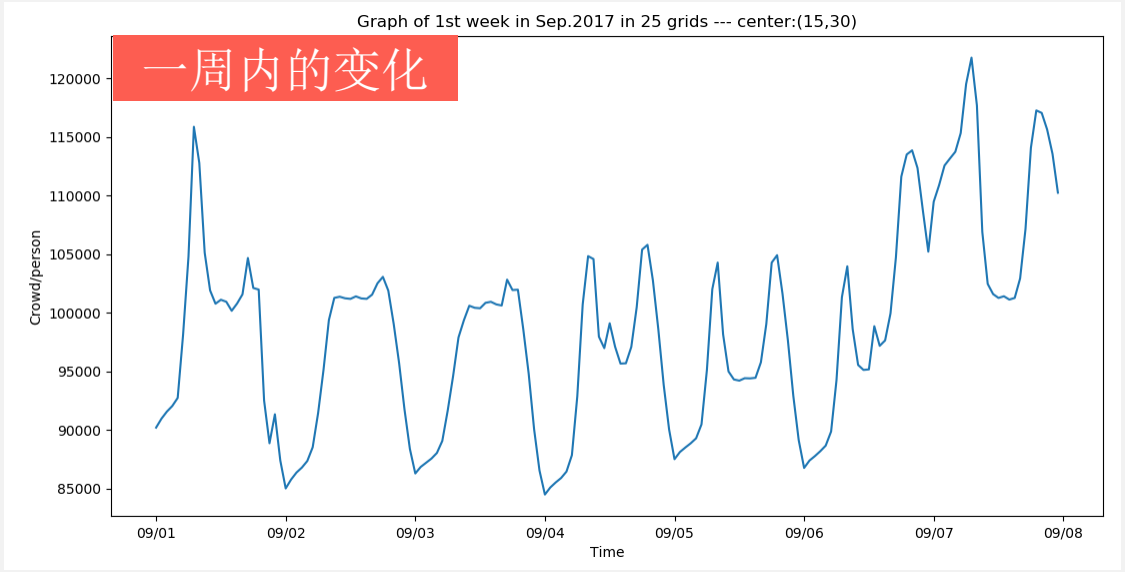
缺失值处理
对数据中的缺失值进行补充或者取历史均值

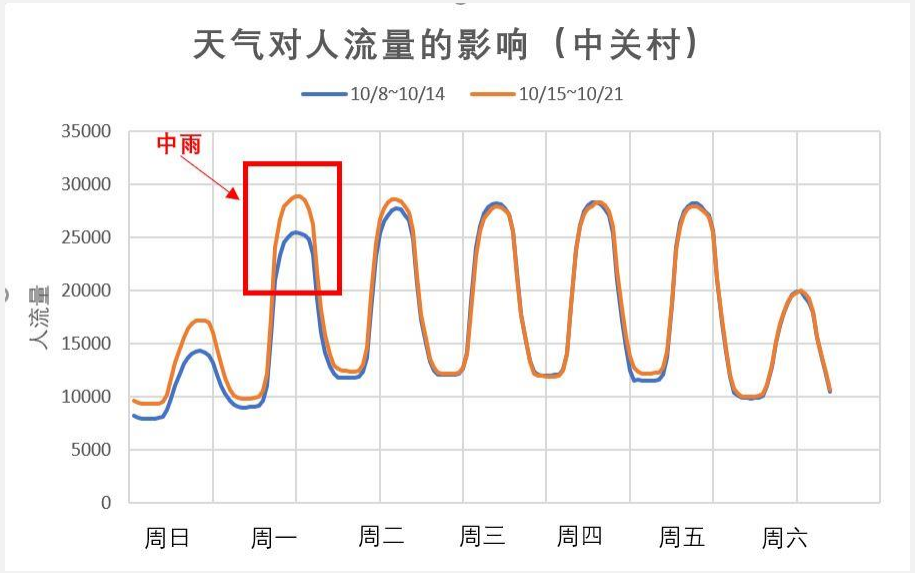
额外数据爬取
对于区域内的北京天气数据进行爬取和处理

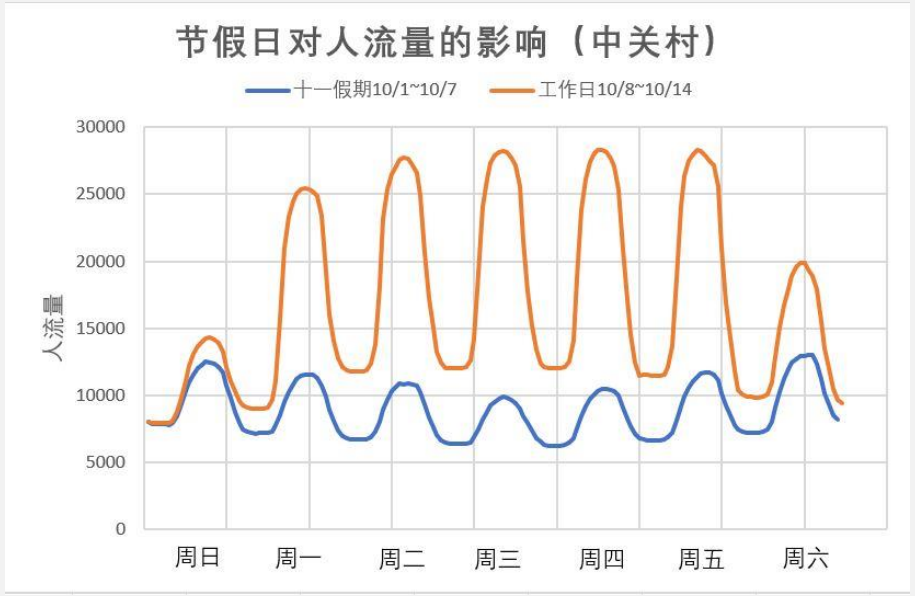
聚类分析
分析人口分布在时间和空间上的集聚性



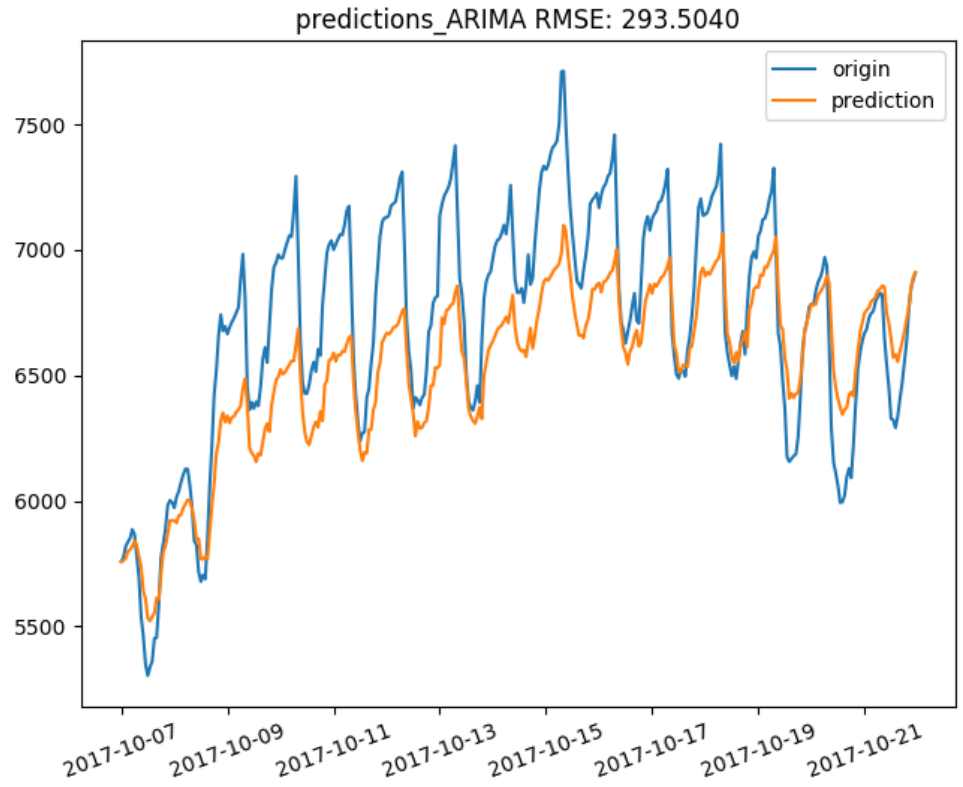
数据预分析：人口流动的周期特性和外部因素影响







ARIMA模型



模型缺陷

采用用于时间序列预测的 AMIMA 模型进行预测发现：可以预测趋势但是存在相当的误差。

分析原因

人口数据的时间分布有特殊的周期规律。即某时刻的数据可能与一定时间前的数据等相关性更高，而不是ARIMA模型中的之前的数据随时间间隔增大相关性逐渐降低。

核心问题

如何发现并考虑人口分布的特殊周期规律？

残差网络人口预测模型

01 影响因素

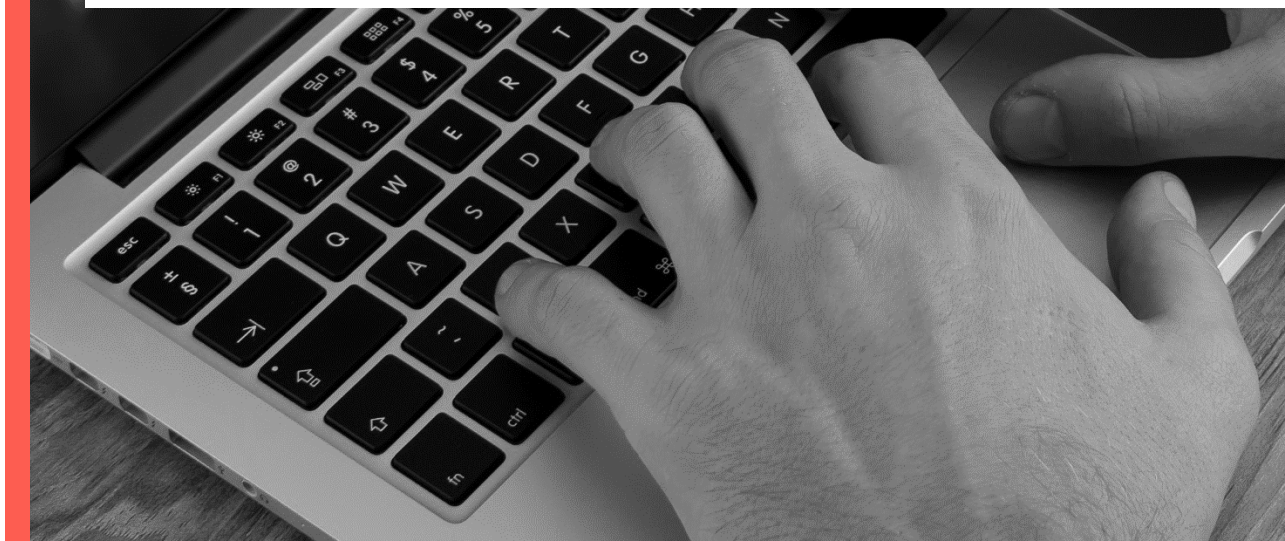
- **在时序上考虑三个时间节点的影响**
  - Xc 邻近时刻（之前的几个小时人口数据）
  - Xp 周期 （昨天的同一时刻前后人口数据）
  - Xq 趋势 （上周的同一时刻的前后人口数据）
- **外部因素影响**
  - 天气和节假日

02 样本划分

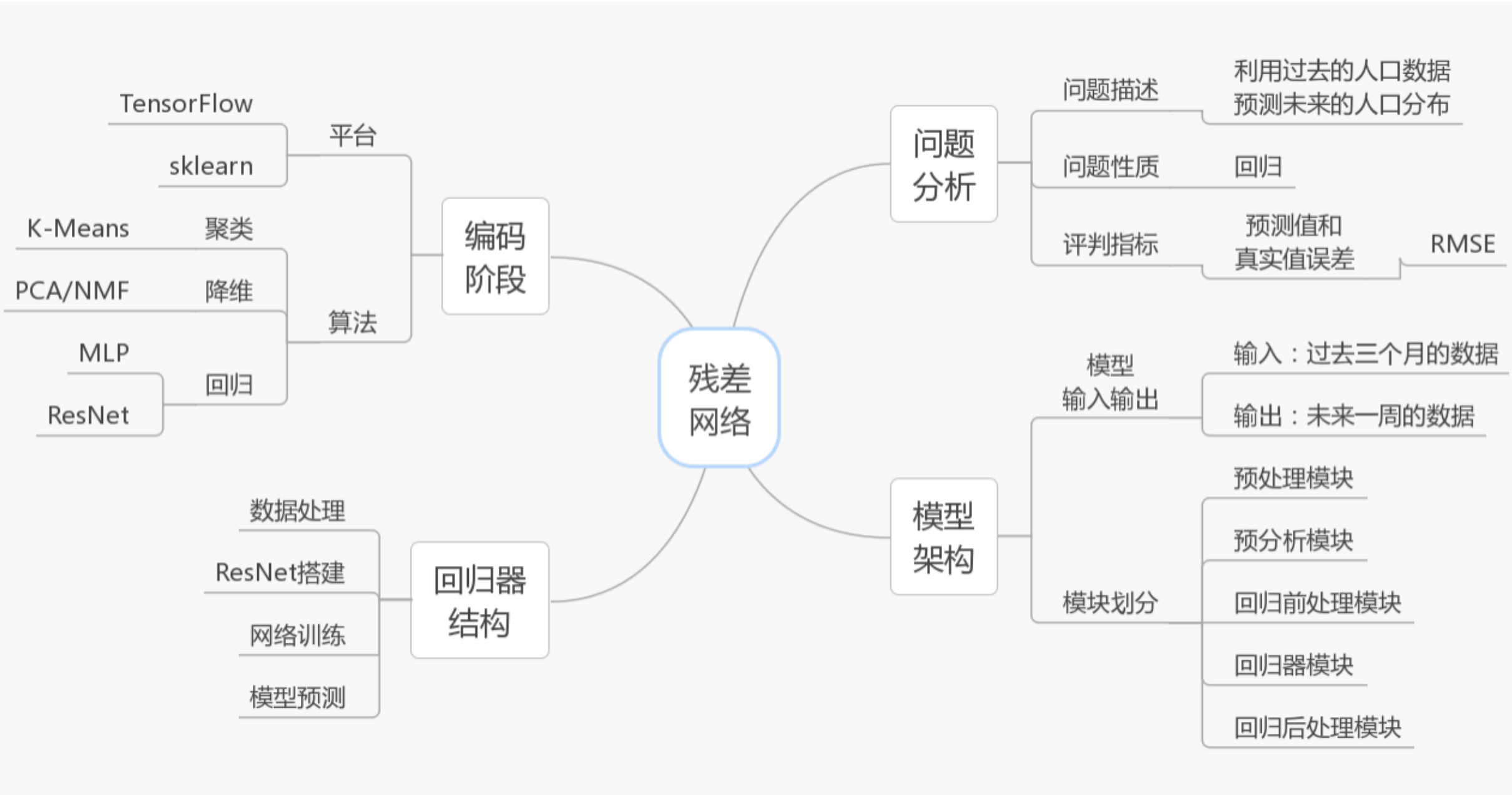
- 以样本中的9/1/2017-11/23/2017的九周为训练集
- 以11/24/2017-12/1/2017的一周为测试集

数据集

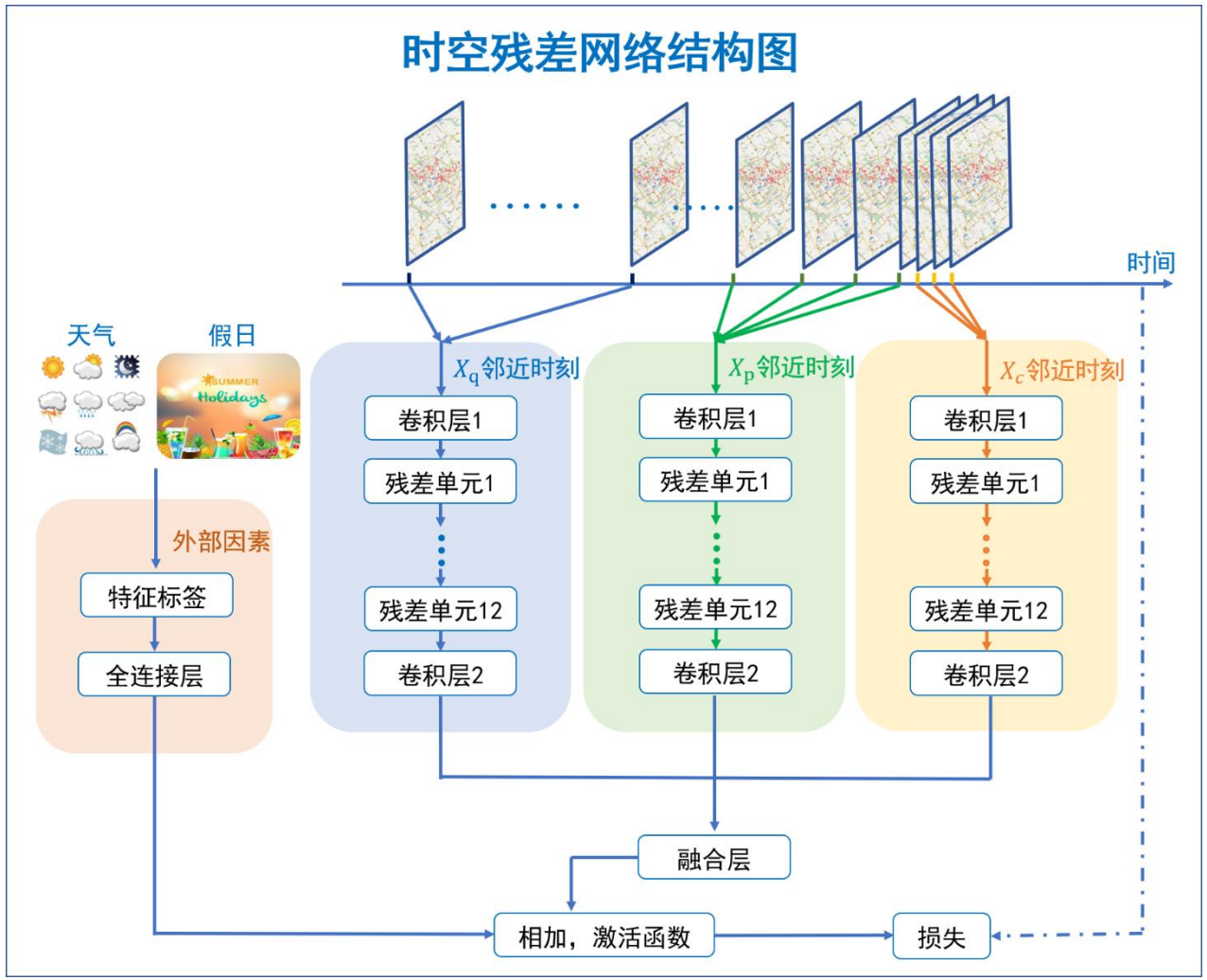
Dataset	PopuBJ
Data Type	Iphone Signal
Location	Beijing
Time Span	9/1/2017-11/30/2017
Time interval	1 hour
Grid map size	(53,54)



人口分布预测模型


残差网络
问题分析
问题描述
利用过去的人口数据预测未来的人口分布
问题性质
回归
评判指标
预测值和真实值误差
RMSE
模型架构
模型输入输出
输入：过去三个月的数据
输出：未来一周的数据
模块划分
预处理模块
预分析模块
回归前处理模块
回归器模块
回归后处理模块
编码阶段
平台
TensorFlow
sklearn
算法
聚类
K-Means
降维
PCA/NMF
回归
MLP
ResNet
回归器结构
数据处理
ResNet搭建
网络训练
模型预测

深度残差网络



Algorithm 1 ResNet Training Algorithm

We first construct the training instances from the original sequence data. Then, RenNet is trained via backprogagation and Adadelta.

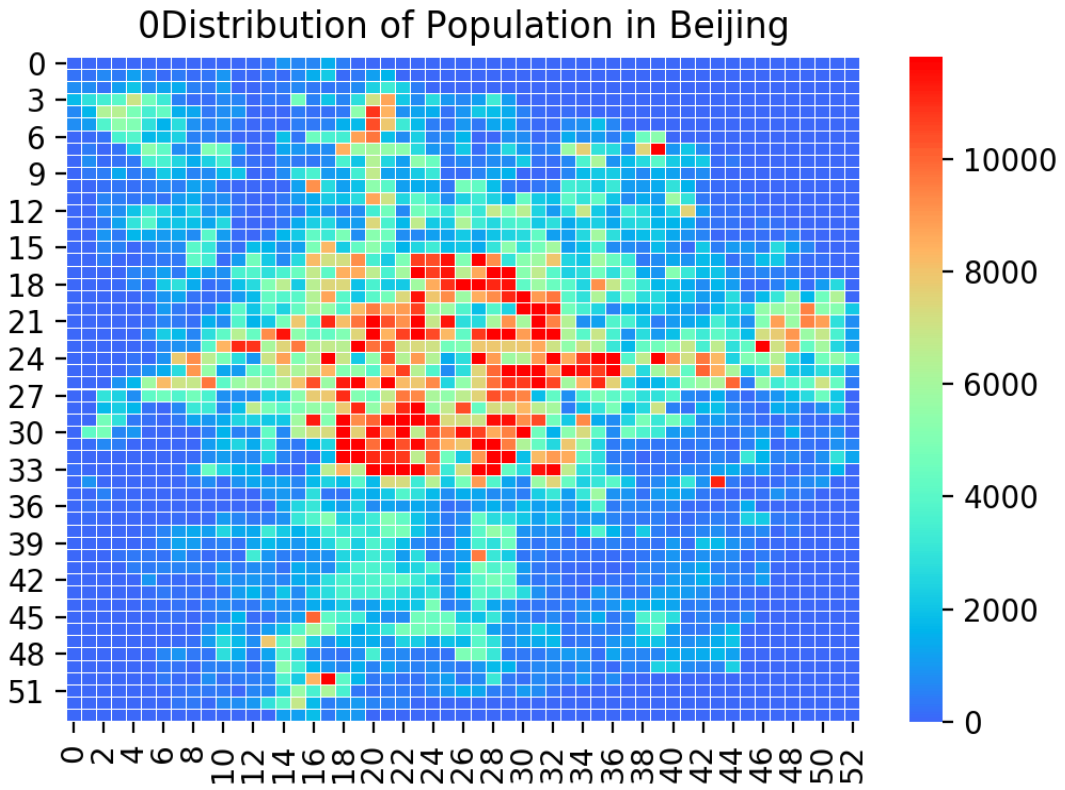
Input: Historical observations: $\{ X_0, \cdots, X_{n-1} \}$
Lengths of closeness, period, trend sequences: l_c,l_p,l_q
Period, trend span: q
Output: learnt ResNet model

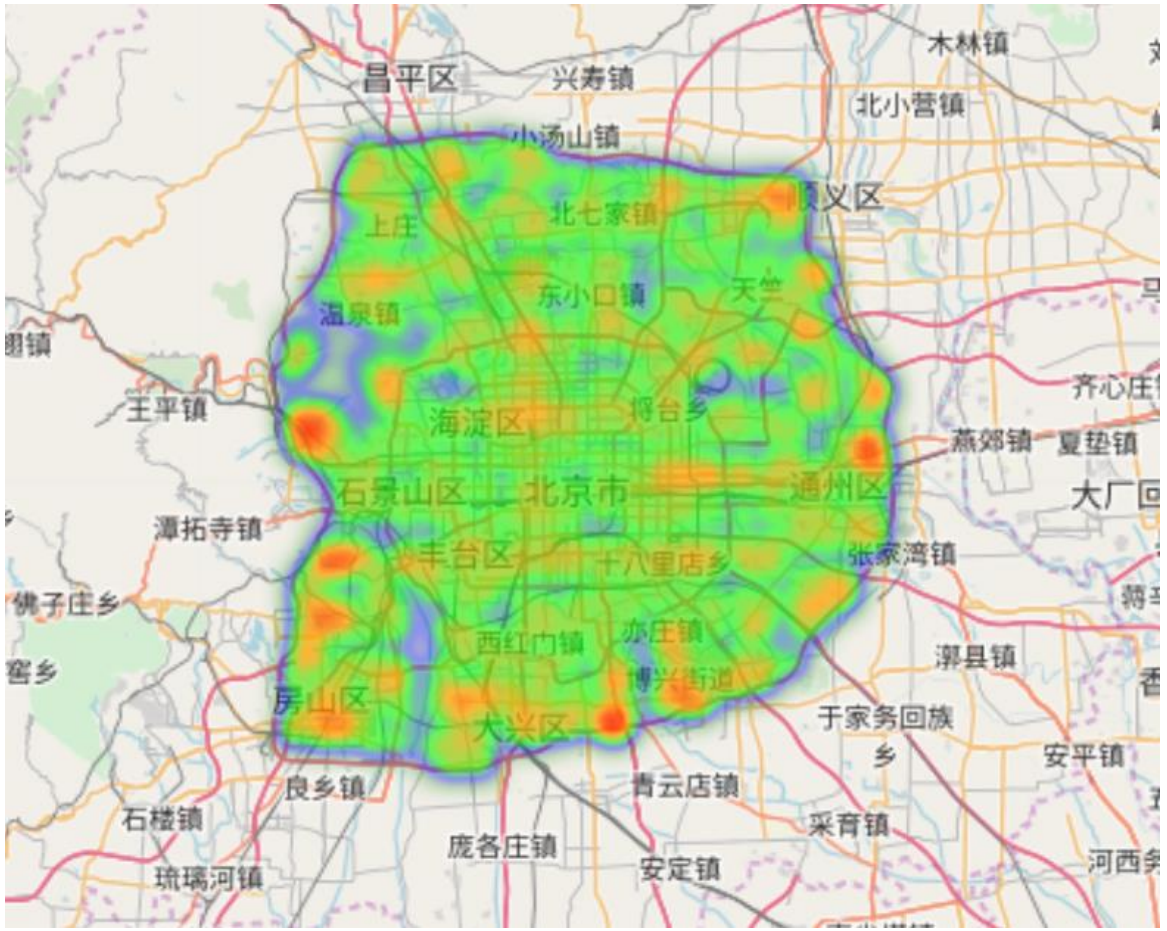
#construct training instances
D <- None
for t in all available time interval [1,n-1] **do**
    $S_c = [X_{t-l_c}, X_{t-(l_c-1)}, \cdots, X_{t-1}]$
    $S_p = [X_{t-l_p*p}, X_{t-(l_p-1)*p}, \cdots, X_{t-p}]$
    $S_q = [X_{t-l_q*q}, X_{t-(l_q-1)*q}, \cdots, X_{t-q}]$
Put an training instance $(\{S_c, S_p, S_q\}, X_t)$ into D
# define the loss
LOSS = MSE(X_t,output of ResNet)
# train the model
# initialize all learnable parameters cita in ResNet
Epoch <- int(numbers)
for epoch in range(Epoch) **do**
    Repeat
    Select an instance D_b from D
    θ by minimizing the LOSS with D_b **Until** all elements of D used
# save model every 10 epoches
    if epoch % 10 == 0 **then**
        Save model

3 预测成果分析

Prediction and Results

结果可视化：人口热力图


0Distribution of Population in Beijing
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
10000
8000
6000
4000
2000
0


昌平区
兴寿镇
木林镇
北小营镇
小汤山镇
顺义区
上庄
北七家镇
东小口镇
天竺
温泉镇
王平镇
海淀区
将台乡
齐心庄
燕郊镇
夏垫镇
大厂回
石景山区
北京市
通州区
潭拓寺镇
丰台区
十八里店乡
张家湾镇
佛子庄乡
西红门镇
亦庄镇
漷县镇
房山区
大兴区
博兴街道
于家务回族乡
安平镇
良乡镇
青云店镇
石楼镇
庞各庄镇
采育镇
琉璃河镇
安定镇
河西务

准确率比较

error<50

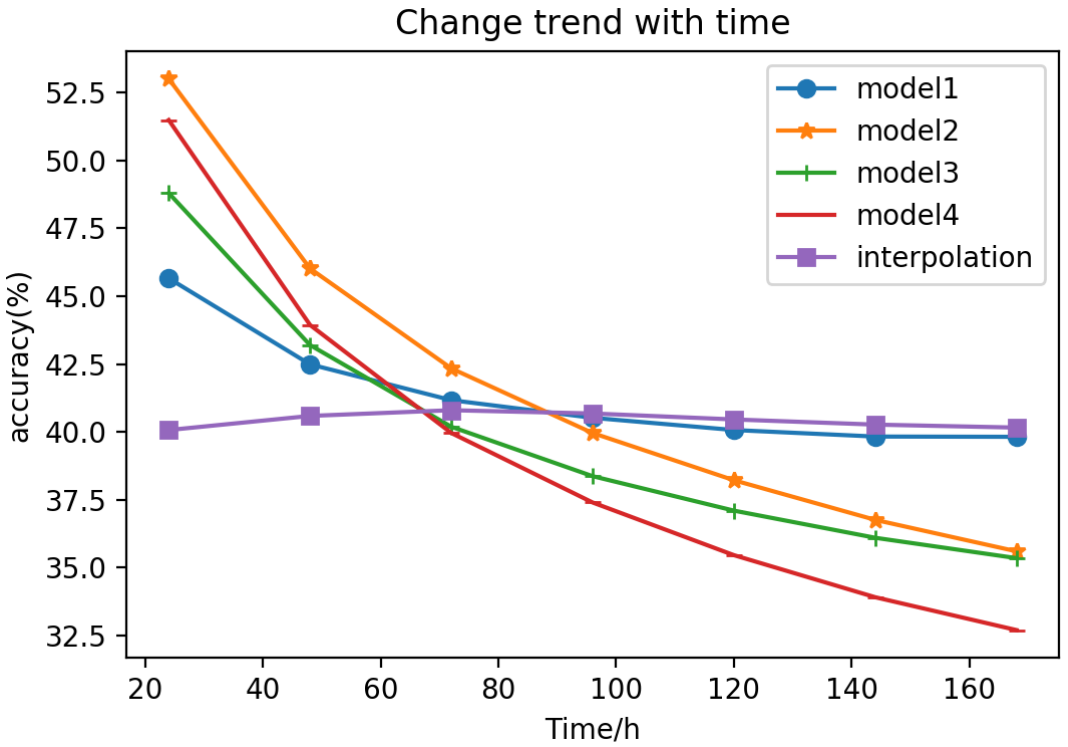
	24h	48h	72h	96h	120h	144h	168h
模型1	45.66%	42.49%	41.17%	40.52%	40.07%	39.83%	39.82%
模型2	53.01%	46.02%	42.34%	39.96%	38.22%	36.76%	35.60%
模型3	48.80%	43.20%	40.19%	38.37%	37.10%	36.10%	35.35%
模型4	51.49%	43.94%	39.96%	37.40%	35.46%	33.91%	32.70%
插值	40.07%	40.59%	40.80%	40.68%	40.46%	40.27%	40.16%

error<100

	24h	48h	72h	96h	120h	144h	168h
模型1	61.11%	57.28%	55.52%	54.71%	54.18%	53.39%	53.39%
模型2	71.79%	62.78%	57.56%	53.90%	51.10%	48.76%	46.84%
模型3	67.35%	59.42%	55.17%	52.37%	50.29%	48.63%	47.36%
模型4	69.63%	59.75%	54.03%	49.98%	46.84%	44.29%	42.23%
插值	50.34%	51.40%	51.73%	51.56%	51.27%	50.99%	50.82%

error<200

	24h	48h	72h	96h	120h	144h	168h
模型1	77.11%	73.98%	72.43%	71.93%	71.46%	71.13%	71.12%
模型2	88.73%	80.81%	75.62%	71.59%	68.24%	65.24%	62.53%
模型3	85.45%	78.03%	73.86%	70.82%	68.38%	66.34%	64.63%
模型4	87.59%	78.74%	72.47%	67.51%	63.37%	59.82%	56.70%
插值	63.19%	64.52%	64.87%	64.59%	64.12%	63.74%	63.48%



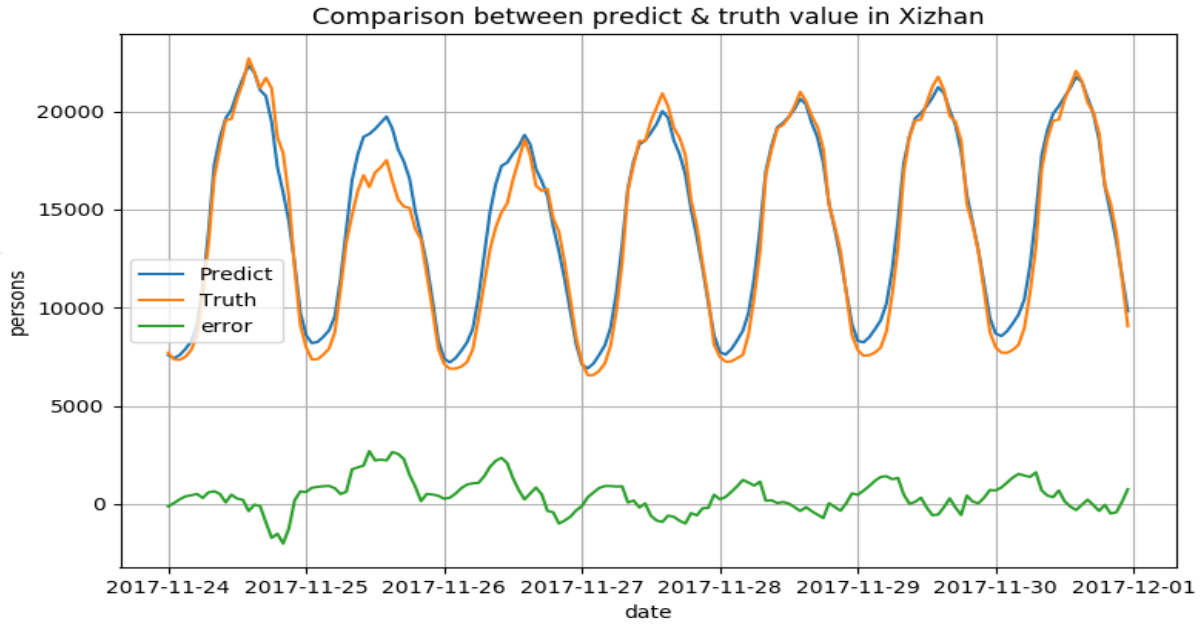
模型预测的准确率都显著高于插值预测模型(baseline)，而不同参数之间的预测准确率也会发生改变

不同模型的预测精度均随着预测时间点的后移而下降：采取逐天分析的方法

区域分析

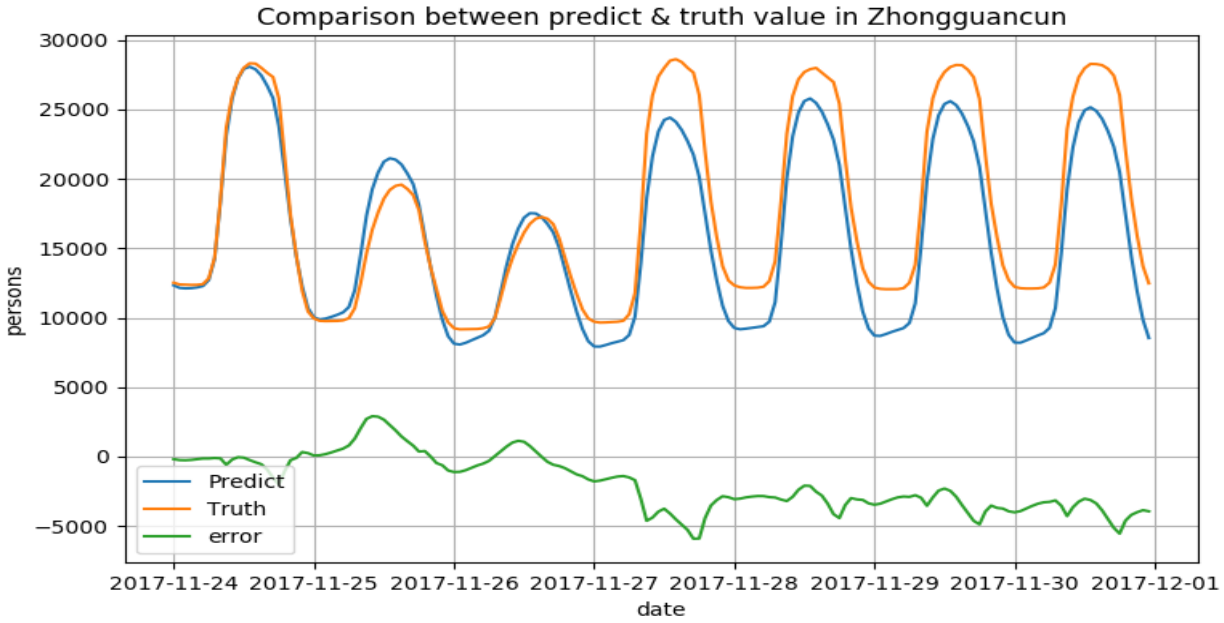
01 北京西站（交通枢纽）

- 预测值和真实值的**符合度比较高**，误差值在前期的分布没有明显的趋势性，表明造成预测误差的一个重要原因可能是随机因素的影响。
- 对于**高峰的预测**是相对比较准确的，可以看出本模型可以在诸如火车站、地铁站等的人流预测和预警中发挥较好的作用。



02 中关村（办公区）

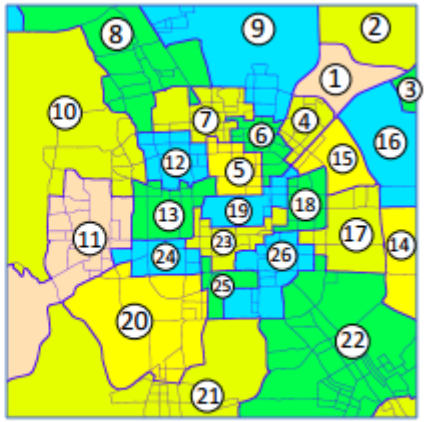
- 预测结果与真值的符合程度很高，并且很好的预测出**节假日对于人口分布的影响**。



模型特点和应用场景展望

- 系统性的考虑了人口流动的时空特性
- 可扩展性：进一步考量节假日等因素
- 可迁移性：模型支持迁移学习，可以在新的应用场景如上海市进行应用
- 展望：可以考虑进一步考虑对于区域的重新划分

(a) Road-based map segmentation



(b) Region grouping using graph clustering

参考文献

[1] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 5(3):38, 2014.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[3] Minh X Hoang, Yu Zheng, and Ambuj K Singh. Fccf: forecasting citywide crowd flows based on big data. advances in geographic information systems, page 6, 2016.
[4] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. Mining individual life pattern based on location history. pages 1–10, 2009.
[5] Jiangchuan Zheng and Lionel M Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data.
[6] Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. Inferring gas consumption and pollution emission of vehicles throughout a city.
[7] Yilun Wang, Yu Zheng, and Yexiang Xue. Travel time estimation of a path using sparse trajectories.
[8] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois.
[9] Charu C Aggarwal. Data streams: Models and algorithms (advances in database systems).
[10] Wang-Chien Lee, John Krumm, Ke Deng, Kexin Xie, Kevin Zheng, Xiaofang Zhou, Goce Trajcevski, Chi-Yin Chow, Mohemad F. Mokbel, and Hoyoung Je-ung. Computing with spatial trajectories. Computing with Spatial Trajectories.
[11] Atsuyoshi Nakamura and Naoki Abe. Collaborative filtering using weighted majority prediction algorithms.
[12] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. Siam Review, 51(3):455–500, 2009.
[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Thank You For listening

All the Codes and Documents are Open-Sourced on Github:

https://github.com/BigDataSystemTHU2018/Project-Unicom

基于信令数据的
人口分布预测

误差分析：RSME

RMSE<300

	24h	48h	72h	96h	120h	144h	168h
模型1	58.30%	50.00%	47.22%	43.75%	41.67%	40.27%	39.88%
模型2	100.00%	75.00%	59.72%	44.79%	35.83%	29.86%	25.59%
模型3	100.00%	68.75%	52.77	39.58%	31.66%	26.38%	22.61%
模型4	100.00%	68.75%	47.22	35.41%	28.33%	23.61%	20.23%
插值	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

RMSE<400

	24h	48h	72h	96h	120h	144h	168h
模型1	95.83%	97.90%	98.60%	87.50%	85.83%	86.68%	88.63%
模型2	100.00%	81.25%	86.11%	72.91%	58.33%	48.61%	41.66%
模型3	100.00%	79.16%	73.61%	62.50%	50.00%	41.66%	35.71%
模型4	100.00%	79.16%	65.27%	48.95%	39.16%	32.63%	27.97%
插值	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

RMSE<500

	24h	48h	72h	96h	120h	144h	168h
模型1	100.00%	100.00%	100.00%	97.91%	98.33%	98.61%	98.88%
模型2	100.00%	100.00%	100.00%	97.91%	85.83%	71.52%	61.30%
模型3	100.00%	87.50%	91.66%	87.50%	78.33%	65.27%	55.95%
模型4	100.00%	95.83%	97.22%	85.41%	68.33%	56.94%	48.80%
插值	20.83%	27.08%	29.16%	28.12%	22.50%	18.75%	16.07%

回归器结构

回归系统的回归器分为4个部分，分别是：数据处理、Resnet搭建、网络训练、模型预测,其代码见 `https://github.com/BigDataSystemTHU2018/Project-Unicom/tree/master/Code/Processor/ResNet`)

模块的主要功能为：

1. 从本地目录获取所有清洗好的数据，将数据按照Resnet的要求加载到内存。
2. 对数据自动分为训练数据和预测数据，其中训练数据喂入Resnet进行模型训练。
3. 对训练好的模型进行预测结果的输出。
4. 支持断点续训，训练过程中模型会每10轮保存一次，可以自由跳转到已经保存好的模型进行继续训练或者模型输出。
5. 支持迁移学习。即利用已经训练好的模型训练新的场景，提高训练的效率和降低对样本数量的需求。
6. getdataTHU.py: 搜索当前目录下所有.csv,.txt文件，以这些文件作为所需的数据。请将待训练的数据放入当前目录。
7. ResnetTHU.py: 定义网路结构，可以根据需要改变卷积核大小(CONV_SIZE)和激活函数。
8. trainTHU.py: 训练模块，也为反向传播模块，根据需要可以改变图片大小 IMAGE_WIDTH,IMAGE_HIGHT,改变通道数NUM_CHANNELS1, NUM_CHANNELS2,NUM_CHANNELS3,同时BATCH_SIZE2,BATCH_SIZE3 ,BATCH_SIZE4也要相应的改变来保持一致（现为取最近3小时，一天前的4小时，一周前的2小时）。
9. predictTHU.py: 预测模块，需要设置预测天数（现为7天）。