项目提供方：中国联通智慧足迹数据科技有限公司

基于信令数据的人口时空分布预测模型构建

第九组：肖飞宇（组长）、牛苒、荆科、沈磊、韦承志、赵嘉欣、张玉生、李司棋

项目中期报告目录

目标和需求

项目需求

通过智慧足迹公司所收集的北京六环以内 1KM x 1KM 网格尺度，分天分小时手机信令数据，预测未来一周相应网格下每天每小时的人口分布

需求解读

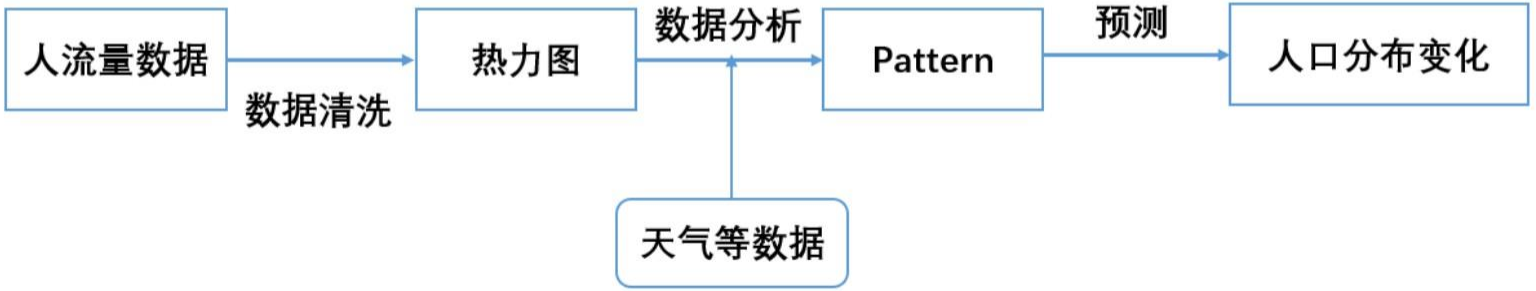
将城市分割成均匀网格，基于交通、气象、时间和事件等多源信息，来综合预测未来每个网格的进入和流出人流数，以便提前启动预警机制，及早疏导人群和车流，保障区域内短时人口密度在安全范围内，从而防范重大交通事故和灾难性城市安全事件（如踩踏）的发生

预期成果

I.
构建基于手机信令数据的人口空间和时间分布精细化预测模型

II.
得到人口的流动以监测和预警

项目分析思路



技术难点

1 信令数据的采样性和稀疏性

某些属性在这个采样上的分布跟它在整体数据上的分布有很大差异,而且通信故障或者传感器故障都会导致数据缺失的问题

2 时空数据的高度复杂性

包括时间属性和空间属性。不同于图像数据和文本数据等，这种特别的属性就意味着传统的深度学习方法不能直接应用

3 人口流动影响因素众多、互相耦合

影响人口流动的因素可能包括：与区域里面前一个小时有多少人进和出有关系；与周边区域有多少人进和出也有关系；甚至很远的地方有多少人进和出等等……但是这些区域之间的关系又是相互影响的

智慧足迹简介


建立
中国联通集团控股
中国联通和西班牙电信合资
成立的专业大数据公司
Smart Steps
全球领先的位置大数据技术
“线下位置+线上行为”
T”型产品战略赋能产业生态
横向开放合作&纵向垂直应用
数据产品&行业垂直产品
明星服务
城市规划、商企洞察选址、金融位置应用
合作与奖项
今

数据分析市场分析

智慧足迹 SmartSteps

清华同衡 T-H-U-P-D-i

TrafficData 晶众股份

智慧足迹

位置大数据应用服务商

上海数慧

大数据在城市规划应用服务商

清华同衡

规划规划设计院
规划研究与数据分析结合

晶众股份

大数据交通应用服务商

案例与应用场景

营销洞察

客流选址

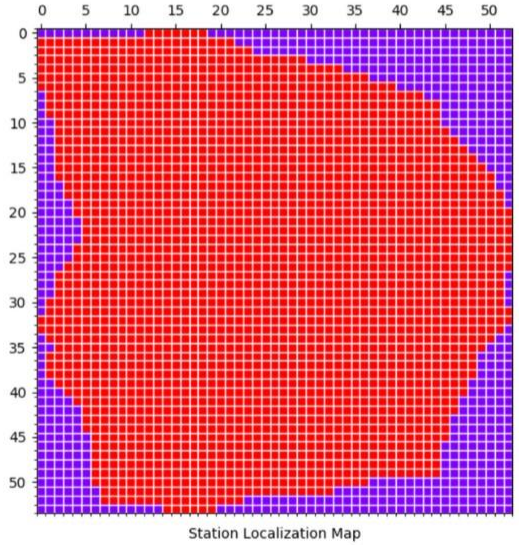
金融位置

广告精准投放

智慧城市规划

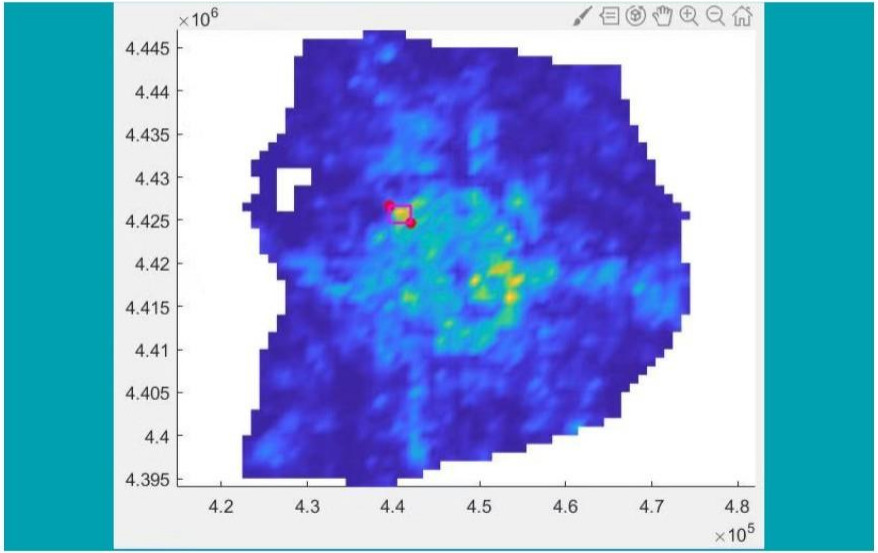
数据预处理

确定基站编号与地理位置关系 1/2



基站所覆盖区域被基站划分为54行53列的网格

且基站编号从网格左上角开始从左到右从上到下依次加一

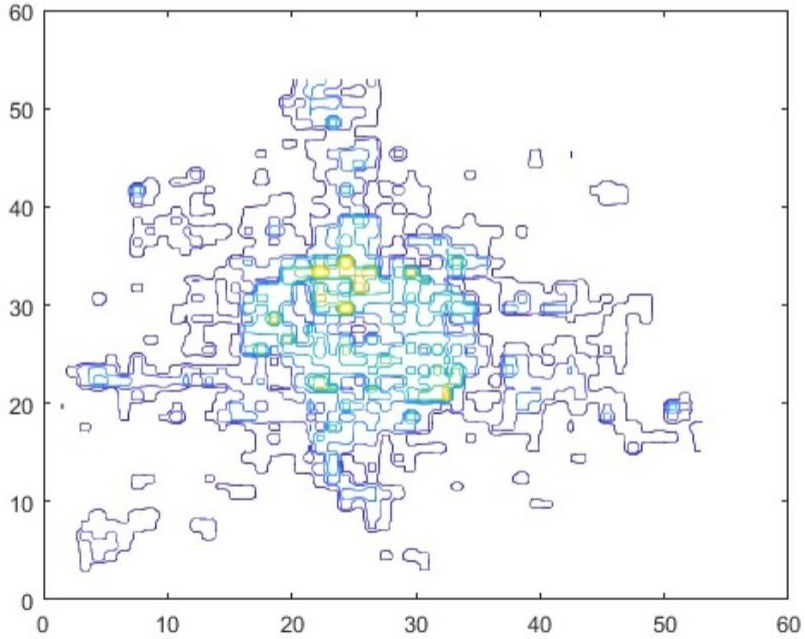


某一格数据缺失情况

使用同一格前后时刻的均值进行补全

数据预处理

确定基站编号与地理位置关系 2/2



原因

每个网格的人口数据不仅是时序的，还与空间分布有关

例如某网格人数的增加与周围网格人数减少相关。前期的数据预处理得到网格分布可能为后续预测提供方便

原始数据

基站编号---wkt坐标对应形式

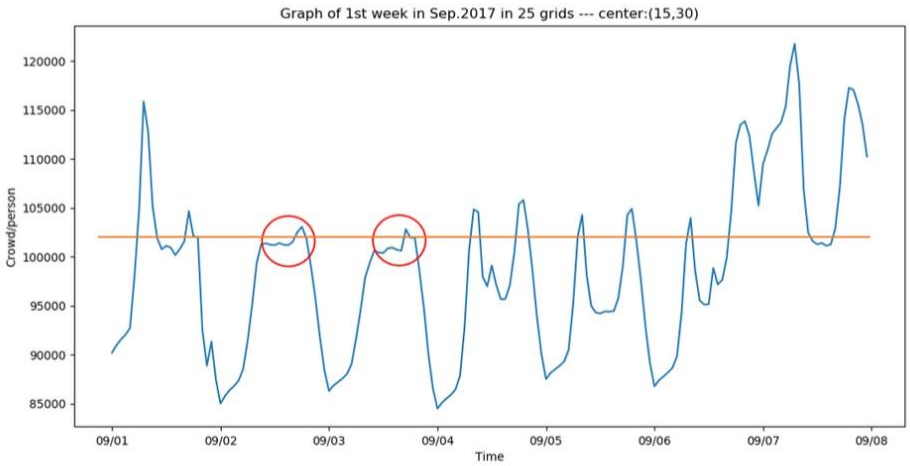
观察

i 基站编号不是从零开始连续的，说明有些位置如山、河等没有人流量数据

ii 只给出了基站编号对应的wkt坐标数据，需要处理得到基站编号基站位置分布的关系

短周期规律观察

i 观察以(15,30)为中心25个网格的一周内人口数量变化



- **说明**

  取25个网格观察，是希望弱化网格间人口流量的影响，主要观察时序变化规律

- **观察**

  i

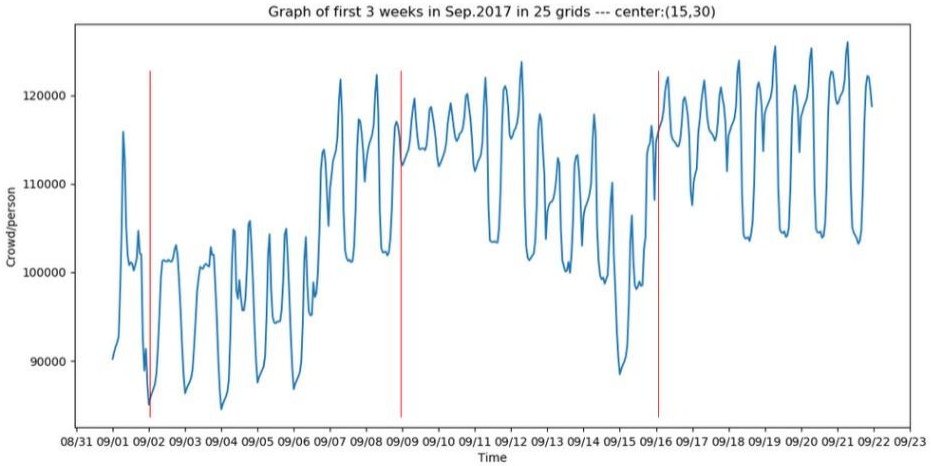
  每天的人口驻留数量有明显的周期性早晚高峰变化。凌晨左右的人口数量最低

  ii

  已知09/02和09/03是周末，发现周末的人口数据与周中人口数据有所不同

长周期规律观察

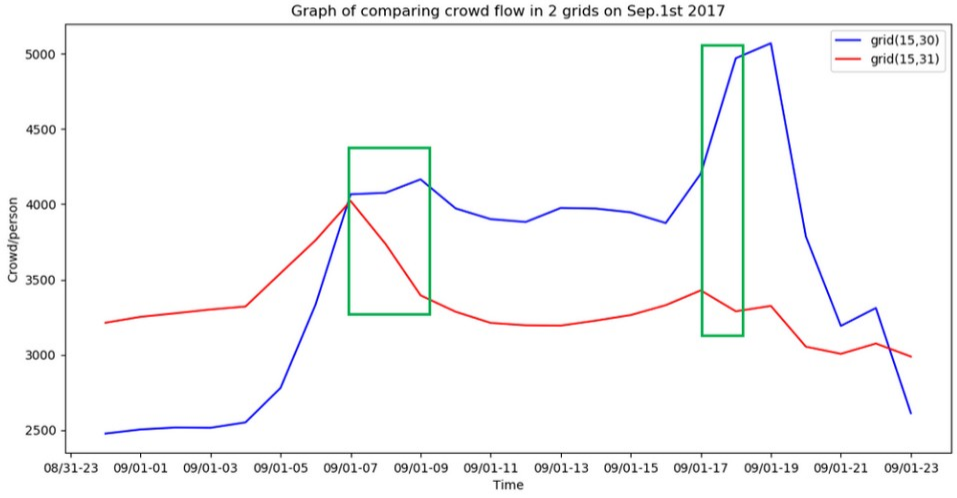
ii 观察以(15,30)为中心25个网格的三周内人口数量变化



- **观察**
  - 可以看到人口数量在各周间也有所波动
  - 周期和通常意义的星期概念不吻合

相邻区域关联性观察

iii 比较(15,30)(15,31)两个网格一天的人口流动



- **观察**

i

这是两个相邻网格的人口数据在一天的变化

ii

观察绿色框，可以发现一个网格人口增多另一个网格人口减少，其原因可能与人们在这两地移动相关

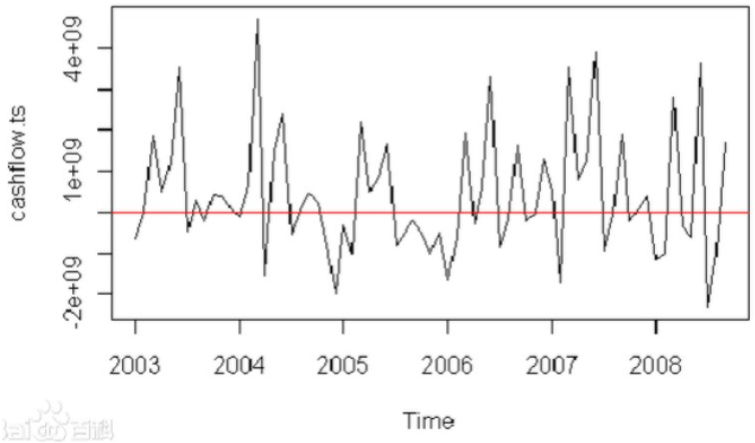
iii

这在一定程度上体现网格位置对人口流动的影响

预测方法（案例）

方法一 ARIMA (Autoregressive Integrated Moving Average) 时序数据分析模型

ARIMA模型示意图

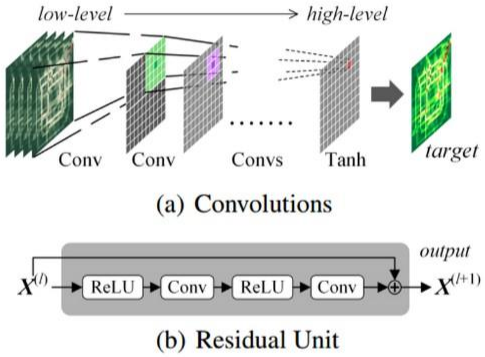


- **对于非平稳的时序序列，常采用差分自回归移动平均(ARIMA)模型进行分析**
- **分析**

  这种方法适用于对一些感兴趣的网格进行初步求解

预测方法（案例）

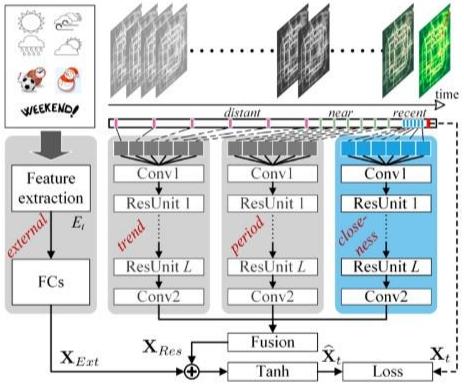
方法二 时空残差网络DeepST方法



i 时序部分

考虑三个因素不同长度周期因素

之前的几个小时/天同一时刻/周同一时刻



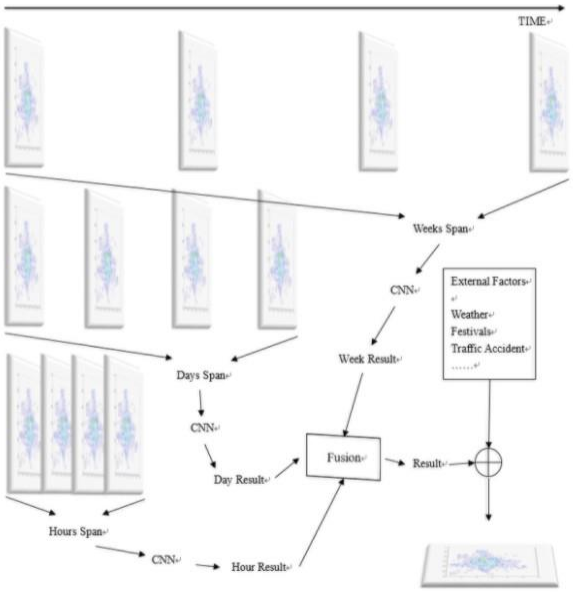
ii 加入外部因素的整体模型

加入天气因素的考量

Zhang J, Zheng Y, Qi D. **Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction**[J]. 2016.

V 技术方案

项目技术实现方案



- **两种信令数据向量化方法&输入输出选择**

  p=f(id,time,week(weather,aircondition,...))

  p=f(id,time,week,p(lasthour),p(yesterday),p(lastweek),p(lastmonth))

  两个模型的差异 是否考虑显性的历史因素

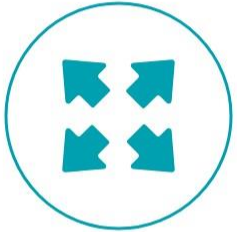
  其中，p代表人口数据，可以通过维度的不同进行不同的表达：一维表示驻留人数， 三维表示驻留人数，出发人数和到达人数

  采用有监督人工神经网络，期望输出为驻留人数、出发人数、到达人数，其基本架构如图所示。

  利用期望输出与网络输出之间的误差建立学习信号反向传播，修正网络权重。

  输入为数据的特征，人工选取：星期几/具体小时/位置/天气/前一小时的人数。同时我们以获得数据中的驻留人数、出发人数、到达人数作为标签，期望输出。

两种实现方式



方案A：多输入多输出
（针对所有点的神经网络模型）

类似于图片处理，2265个基站
输入即为2265*n，其中n为输入特征（星期/小时/ID（位置特征）等），
输出为2265*3，3为标签个数（驻留人数/出发人数/到达人数）

方案B：少输入少输出
（针对某一点的神经网络模型）

针对某一个基站或者某一块区域进行训练
输入为特征个数(星期/小时/ID等)，
输出为3，即驻留人数、出发人数、到达人数

训练出来的模型是针对某一个基站或者某一块区域
（9个基站左右，表示9KM^2的一个区域）的人口情况
做分析时候要训练多点得到一个区域的数据进行分析

异构数据源的融合

在项目进行过程中，我们需要将来自不同数据源的数据融合到一起。例如，项目方提供的数据、天气数据、交通数据等信息。如何将不同数据源的数据高效的融合到一起，我们尝试用以下方法：

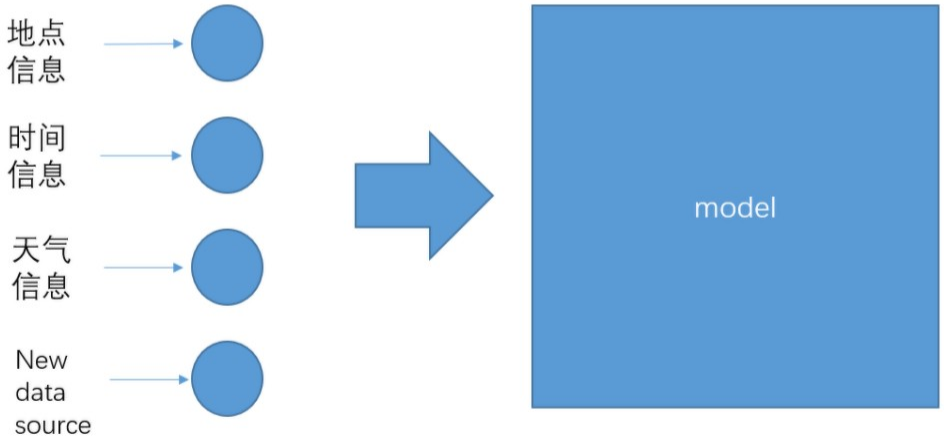
方法一：将不同数据源的数据放到同一个特征向量作为输出

方法二：在不同阶段使用不同数据源的数据

方法三：将不同数据源的数据喂入模型的不同部分

异构数据源的融合

方法一：将不同数据源的数据放到同一个特征向量作为输出

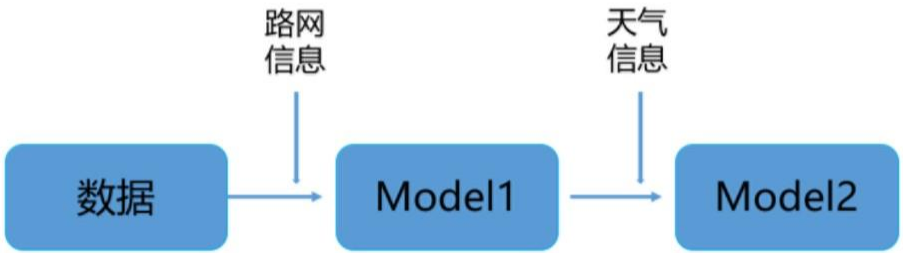


- **将不同数据源的数据提取到的特征同等看待，一同放入模型的输入向量**

E.G.
模型的一维放入地点信息、一维放入时间信息、一维放入天气信息等等

异构数据源的融合

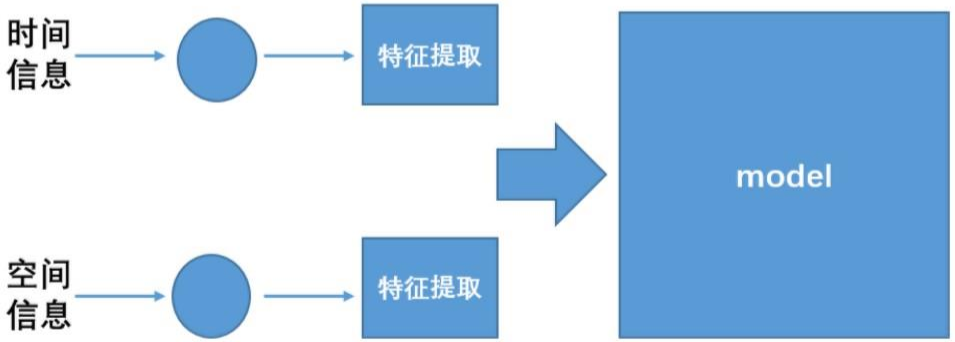
方法二：在不同阶段使用不同数据源的数据



- 先用路网信息或者行政区域划分将区域变为多个子区域，然后再在每个子区域里面用其他信息例如信令数据、天气等信息

异构数据源的融合

方法三：将不同数据源的数据喂入模型的不同部分



- 将时间信息放入时间分类器模型，空间信息放入空间分类器模型，输出的结果再一同放入模型学习

VII 项目进度

项目时间安排