**北京邮电大学**

**本科毕业设计论文**



**名称：基于腾讯定位数据的异常事件检测算法**

**学院： 信息与通信工程学院**

**班级： 2014211112**

**姓名： 林文鼎**

**学号： 2014210328**

**班内序号： 07**

**指导教师： 别红霞**

**验收日期： 2018年5月7日**

**摘 要**

摘要还没写

**关键词：**关键词1；2；3

目 录

1. 绪论 1

1.1 课题背景 1

1.2 异常检测研究现状 1

1.2.1 基本概念与挑战 1

1.2.2 异常检测算法分类 2

1.3 论文主要工作 3

1.4 论文章节安排 3

2. 异常检测算法基础 3

2.1 时序曲线离群点检测 4

2.1.1 极大似然估计 4

2.1.2 离散序列小波变换 4

2.1.3 差分算法 4

2.1.4 局部异常因子 4

2.2 时序曲线预测 4

3. 定位数据 4

3.1 腾讯定位数据的形式 4

3.2 数据分析及应用 5

3.2.1数据的小时变化规律 5

3.2.2 数据的日变化规律 5

3.2.3 数据的总变化特征 5

3.3 数据预处理及异常分析策略 5

3.3.1 数据预处理 5

3.3.2 台风天异常分析策略 6

4. 基于曲线分析的定位数据异常检测 6

4.1 基于小波的异常检测算法 6

4.1.1 离散序列小波变换 6

4.1.2 结果分析 7

4.1.3 改进算法 7

4.2 基于极大似然估计的异常检测算法 7

4.2.1 极大似然估计与3σ准则 7

4.2.2 结果分析 7

4.3 局部异常因子检测算法 8

4.3.1 局部异常密度 8

4.3.2 结果分析 8

4.4 基于差分的异常检测算法 8

4.4.1 差分算法 8

4.4.2 结果分析 8

4.6 曲线异常检测算法总结 8

5. 定位数据的预测分析 9

5.1 异常的检测与预测 9

5.2 基于动态神经网络的定位数据预测 9

5.2.1 动态神经网络 9

5.2.2 定位数据预测 9

6. 基于区域的定位数据异常检测 10

6.1 基于图像的异常区域检测 10

6.1.1 相邻帧间差分法 10

6.1.2 高浮动区域检测 10

7. 总结与展望 11

7.1 内容总结 11

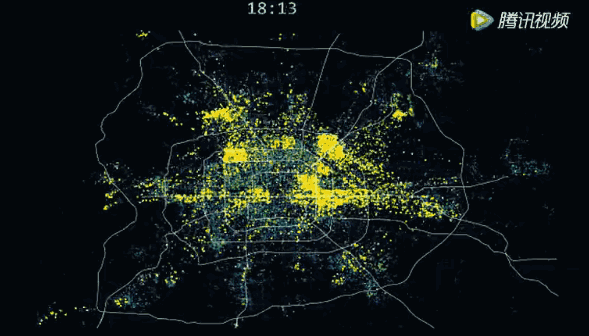
7.2 未来展望 11

1. 绪论

本章主要介绍了定位数据的异常事件检测的课题背景及其研究意义，其次介绍了异常检测这一领域的基本概念及常用算法分类，最后对论文的主要研究工作进行了总结并阐述了本文的行文章节安排。

1.1 课题背景

随着GPS定位，传感器网络和高速无线通信等技术的日益发展，越来越多的终端定位数据被收集和保存在应用服务器，它们是各地人口密度的一个衡量依据。除了定位数据本身所体现的人口密度空间特征，在相同的空间位置不同的时间点上进行记录还可以得到定位数据的人口密度时序特征。而通过分析某片区域上的时序定位数据，可以得到该区域上人口密度的变化特征，例如从北京市每日的滴滴打车的定位数据可以明显地总结出以下特征：滴滴车辆在早高峰时将大量住在郊区的人群运送至各大工作区（例如中关村和国贸区域），而在晚高峰时它们又将人群从工作区运送回家，如下图所示。这种区域性的时序定位数据反应出了北京市的日人口密度变化特征。

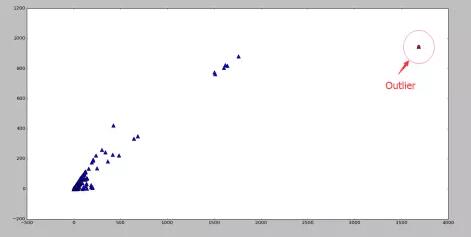
 

在这种区域性的时间序列中，也会在某些时间点上出现这样的观测点，它们较以往同时段的数据，例如某个星期天与前几周的星期天来说，会有一个明显的波动，这就是时序数据点中的异常值。分析这些异常点也是一个很重要的课题，通常对于区域性时序特征模型的建立，这些异常点是应当被剔除的噪声，它们会对模型的预测功能产生极大的阻碍。但同时异常点也可以作为一些突发事件（如异常气象，交通管制等）的判断因素，在异常事件发生时，区域的时序定位数据会在某一个时间间隔中出现较大的落差，即异常波动，通过异常检测算法将上述波动检出并分析，可以使有关部门察觉到异常状况的发生并及时做好应急响应。

1.2 异常检测研究现状

1.2.1 基本概念与挑战

“异常”是指数据特征不符合该特征一般所隶属区间的现象。寻找异常是一个非常困难的课题，其难点主要来源于以下两个角度：首先，“异常”通常情况下只是一个定性的概念，偏离正常数据多少可以被界定为异常没有一个定量的比例数值，那么对于那些处于异常非异常边界线附近的异常数据来说，完全可以把边界线略微移动，使其能被归类为正常的数据；再者，用于划定数据特征正常区间的正常样本中有时也会存在异常数据，导致划定边界线偏差或是训练出的预测模型不准确。同时，考虑到正常的数据量远大于异常数据，使用机器学习的方法进行训练时很容易使网络结构偏向于正常数据的分布，即过拟合导致无法检测出异常。



为了便于分析，异常也有多种分类：通常情况下直观理解的异常指的是点异常，其含义是多个数据实体中，如果存在一个实体对于其他实体来说有极大的偏差，那么这个实体数据所对应的特征就表明一种点异常。而另外一种异常被称为环境异常，它与点异常中的异常概念是一样的，表征的是一个数据实体在特定环境中的异常，存在某种限定条件。这种异常类型的数据实体有两部分组成：环境属性与行为属性。环境属性表征了数据实体所处在的环境，例如时间序列数据的时间点，空间数据的地理坐标；行为属性表征了在上述特定环境属性下区分数据实体的属性，类似于地理数据的某地降雨量，行为属性即固定了环境属性后的数据特征，例如已知某地理坐标上的定位数据量。在本课题中，数据属于时序性的定位数据，而其中的异常是一种环境异常。环境属性即是时空坐标与地理坐标，行为属性是在某时间点上某地理坐标下的定位终端数量。

1.2.2 异常检测算法分类

异常检测是找出数据特征严重不同于预期对象的一个检测过程。传统检测异常的方法分为以下几类：基于分类的异常检测方法，基于最近邻的异常检测方法，基于聚类的异常检测方法，基于统计的异常检测方法。

1. 基于分类的异常检测方法：分类是一种从一组已做好标注的数据实例（训练）中学习模型（分类器），然后使用学习模型（测试）将测试实例分类到其中一个类中的方法。基于分类的异常检测算法以类似的两阶段方式生成一个分类模型从而判断数据的特征是否异常：在训练阶段通过使用已进行标记的样本来训练分类器的模型和参数；测试阶段使用分类器将测试实例分类为正常或异常。基于分类的异常检测算法基于以下假设下实现：在给定的数据特征空间中学习可以区分正常类和异常类的分类器是可行的。

2. 基于最近邻的异常检测方法：近邻分析的概念已用于多种异常检测算法，这些算法都基于以下关键假设：正常的数据实例发生在密集的邻域中，而异常发生在离它们最近的邻居很远的地方。最近邻的异常检测算法需要两个数据实例之间所定义的距离或相似度衡量，而这些距离又针对不同类别的属性有不同的衡量标准。对于连续的属性，欧几里得距离的效果优秀，可以表征出两个数据间的联系性，但在不同情况下也可以使用其他方法计算。对于分类属性，通常使用简单的匹配系数，但也可以使用更复杂的距离度量。对于多变量数据实例，通常为每个属性计算距离或相似度，然后进行合并。

3. 基于聚类的异常检测方法：聚类用于将类似的数据实例分组到集群中。尽管聚类是无监督式学习，但聚类在半监督式学习中的应用也被最近探讨。尽管聚类和异常检测看起来彼此根本不同，但是已经有几种基于聚类方法被应用于异常检测：

* 第一类基于聚类的算法依赖于以下假设：普通数据实例属于数据中的一个集群，而异常不属于任何集群。基于此假设的技术将已知的基于聚类的算法应用于数据集，并将任何不属于任何聚类的数据实例声明为异常。
* 第二类基于聚类的技术依赖于以下假设：正常的数据实例靠近它们最接近的集群质心，而异常距离它们最近的集群质心很远。基于这种假设的技术由两个步骤组成。在第一步中，数据使用聚类算法进行聚类。 在第二步中，对于每个数据实例，计算它到最近的集群质心的距离作为其异常分数。
* 如果数据中的异常自身形成聚类，这些技术将无法检测到这种异常。为了解决这个问题，已经提出了第三类基于聚类的算法，它依赖于以下假设：普通数据实例属于大型且密集的集群，而异常集群属于小型集群或稀疏集群。基于此假设的技术将属于大小和/或密度低于阈值的集群的实例声明为异常。请注意，如果数据中的异常自身形成集群，则这些技术将无法检测到这种异常。

4. 基于统计的方法：统计异常检测技术基于以下关键假设：正态数据实例出现在随机模型的高概率区域，而异常发生在随机模型的低概率区域。统计技术将一个统计模型（通常用于正常行为）与给定数据相匹配，然后应用统计推断测试来确定一个看不见的实例是否与该模型相符。基于应用的测试统计信息从学习模型中生成概率较低的实例被声明为异常。

1.3 论文主要工作

本文基于以上课题背景以及研究现状，基于腾讯地图所提供的时序定位终端地图数据，在已知某一天为异常天（台风过境）的前提下，研究并实现检测出该天为异常天的算法。并且，在同时实现了根据现有时序定位数据预测后几日定位数据的生成模型，最后根据检测结果对区域的定位数据进行了讨论并将上述成果以MATLAB应用的形式输出。具体实现内容包含以下几部分：

1. 对腾讯定位数据进行解析及预处理。首先，由于研究的异常为台风过境时某区域的定位数异常，将定位数据的区域统一标定在该地域的经纬度；其次，对定位数据进一步作图分析，观察在相同位置处定位终端数量一天内的变化、每天同时段的变化，进而确定分析策略；最后，根据数据特点进行预处理便于分析。
2. 对经过处理后的数据基于曲线分析进行异常检测。采用了诸如小波变换，极大似然估计法，差分分析法等传统方法以及一些混合改进算法，对这些算法的效果进行对比分析，总结在该定位数据下哪一种算法实现结果最佳。
3. 根据处理后的数据以及曲线异常检测的结论，设计并实现曲线预测模型。采用动态神经网络使用现有数据进行训练，并使用后续补充的数据进行验证，提出使用预测的方法可以解决不确定异常天数的结论。
4. 在前几章中经过数据的处理及算法的设计，异常天能够被成功检出并且根据现有的数据可以实时判断异常。在这一章中我们使用视频分析的帧间差分法以及滑动窗的方法确定造成整体异常的区域。（使用未知的国庆天异常来判明）
5. （应用）

1.4 论文章节安排

第一章 绪论。本章主要介绍了定位数据的异常事件检测的课题背景及其研究意义，其次介绍了异常检测这一领域的基本概念及常用算法分类，最后对论文的主要研究工作进行了总结并阐述了本文的行文章节安排。

第二章 异常检测算法基础。本章主要介绍了异常检测的算法理论基础，涵盖后续几章所用的主要算法的一些理论背景，包括了极大似然估计，小波变换，神经网络等相关概念。

第三章 定位数据。本章主要介绍了本课题所研究的腾讯定位数据的形式，并在MATLAB中作图分析其数据特征，包含时序特征及地理特征，最后根据分析结果讨论如何进行检测。

第四章 基于曲线分析的定位数据异常检测。 本章主要基于第三章数据分析的结果从曲线的角度对数据进行异常检测，研究了在数据中找出异常的方法。使用了例如小波变换、极大似然估计和邻域的方法进行了分析，并对这些方法的效果进行了讨论。

第五章 定位数据的预测分析。 本章主要对异常检测的另一种思路进行了探讨，即实时判断数据是否异常的检测方法。采用了动态神经网络训练时序数据预测模型，并根据预测与实际值比对的结果对异常检测的效果进行了分析。

第六章 基于区域的定位数据异常检测。 本章基于前述两章的异常检测结果，检测造成异常的重点区域。采用了视频分析的策略，例如帧间差分法，将异常的核心区域变化特征凸显并使用滑动窗标记。

第七章 总结与展望。本章主要对基于腾讯定位数据的异常检测算法进行分析总结，针对实验结果进行分析，获得本文方法存在的缺点并且提出解决存在问题的有效方法，提高异常检测的准确率和平台使用的有效性。

1. 异常检测算法基础

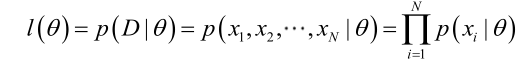
本章主要介绍了本课题所研究的异常检测的算法基础，因为在后续几章中要根据本课题所研究的数据对象对算法进行改写，所以在本章中主要介绍理论基础，后续章节中再具体阐述实际应用策略。

2.1 时序曲线离群点检测

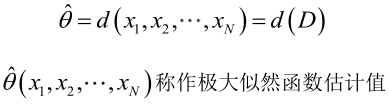
2.1.1 极大似然估计

https://blog.csdn.net/zengxiantao1994/article/details/72787849

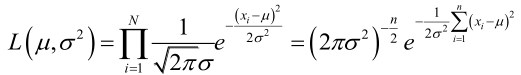
极大似然估计是利用已知的样本信息，观察其结果，利用试验结果去反推最有可能（最大概率）导致这样结果的参数值。极大似然估计提供了一种在给定观察数据的条件下来评估模型参数的方法，即“模型已知，参数未知”的情况。例如已知数据集的采样是独立且为正态分布时，我们可以使用极大似然估计法求出参数μ和σ的值。设已知样本集为：

似然函数（linkehood function）：联合概率密度函数称为相对于的θ的似然函数。

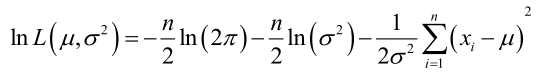
如果是参数空间中能使似然函数https://img-blog.csdn.net/20170528003236220 最大的θ值，则应该是“最可能”的参数值，那么就是θ的极大似然估计量。它是样本集的函数，记作：



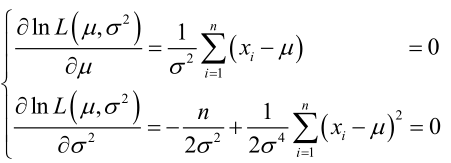
样本服从正态分布https://img-blog.csdn.net/20170528003917176，则似然函数为：



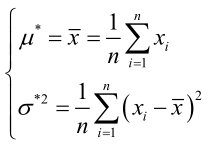
        它的对数：



        求导，得方程组：



        联合解得：



        似然方程有唯一解https://img-blog.csdn.net/20170528004743185：，而且它一定是最大值点，这是因为当https://img-blog.csdn.net/20170528004747290或https://img-blog.csdn.net/20170528004751982时，非负函数https://img-blog.csdn.net/20170528004756951。于是U和https://img-blog.csdn.net/20170528004801165的极大似然估计为https://img-blog.csdn.net/20170528004743185。

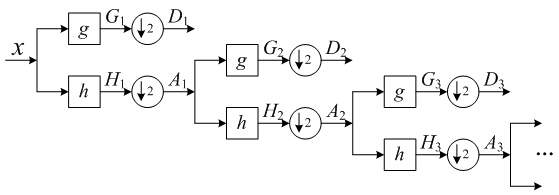
2.1.2 离散序列小波变换

https://blog.csdn.net/jbb0523/article/details/42554693

小波变换（wavelet transform，WT）是一种新的变换分析方法，它继承和发展了短时傅立叶变换局部化的思想，同时又克服了窗口大小不随频率变化等缺点，能够提供一个随频率改变的“时间-频率”窗口，是进行信号时频分析和处理的理想工具。

https://img-blog.csdn.net/20150109143113218

离散序列的小波变换就是基于著名的Mallat算法，实际上Mallat算法也是针对连续信号的，但在每一层的分解过程中（这里说的每一层实际上就是小波中尺度的概念），各层分解系数之间有着某种关系，具体关系如下图所示Mallat算法框图所表示：

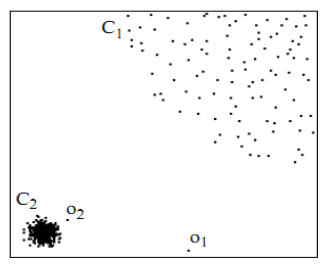


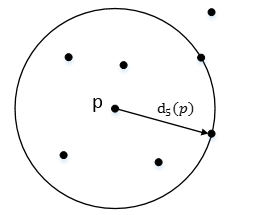
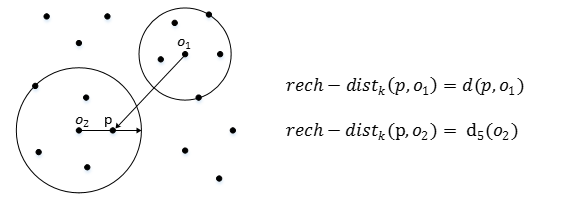
初始系数x与其第一层分解后的高频系数D1（细节部分Detail）的关系是x经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是x经过低通滤波器h滤波后再下采样；然后继续对低频系数A1进行第二层分解，低频系数A1与其第二层分解后的高频系数D2（细节部分Detail）的关系是A1经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是A1经过低通滤波器h滤波后再下采样；后面依次类推即可。由于一直在下采样，所以虽然滤波器系数g和h不变，但其滤波带宽一直在减半。初始系数是怎么来的呢？肯定是根据信号得到的，最简单最粗糙的办法就是对信号直接抽样。这是对连续信号进行正交小波分解，有了这些系数，再利用正交小波基，就可以表示出信号了，这类似于连续周期信号的傅里叶级数分解。

2.1.3 最近邻及邻域算法

<https://blog.csdn.net/wangyibo0201/article/details/51705966>

如图2，对于C1集合的点，整体间距，密度，分散情况较为均匀一致，可以认为是同一簇；对于C2集合的点，同样可认为是一簇。o1、o2点相对孤立，可以认为是异常点或离散点。现在的问题是，如何实现算法的通用性，可以满足C1和C2这种密度分散情况迥异的集合的异常点识别。LOF可以实现我们的目标。



面介绍LOF算法的相关定义：   
　　**1)** d(p,o)d(p,o)：两点p和o之间的距离；   
　　**2) k-distance：第k距离**   
　　　　对于点p的第k距离dk(p)dk(p)定义如下：   
　　　　dk(p)=d(p,o)dk(p)=d(p,o)，并且满足：   
　　　　　　a) 在集合中至少有不包括p在内的kk个点o,∈C{x≠p}o,∈C{x≠p}， 满足d(p,o,)≤d(p,o)d(p,o,)≤d(p,o) ；   
　　　　　　b) 在集合中最多有不包括p在内的k−1k−1个点o,∈C{x≠p}o,∈C{x≠p}，满足d(p,o,)<d(p,o)d(p,o,)<d(p,o) ；   
　　　　p的第k距离，也就是距离p第k远的点的距离，不包括p，如图3。   
　　　　   
　　**3) k-distance neighborhood of p：第k距离邻域**   
　　　　点p的第k距离邻域Nk(p)Nk(p)，就是p的第k距离即以内的所有点，包括第k距离。   
　　　　因此p的第k邻域点的个数 |Nk(p)|≥k|Nk(p)|≥k。   
　　**4) reach-distance：可达距离**   
　　　　点o到点p的第k可达距离定义为：   
　　　　reach−distancek(p,o)=max{k−distance(o),d(p,o)}reach−distancek(p,o)=max{k−distance(o),d(p,o)}   
　　　　也就是，点o到点p的第k可达距离，**至少**是o的第k距离，或者为o、p间的真实距离。   
　　　　这也意味着，离点o最近的k个点，o到它们的可达距离被认为相等，且都等于dk(o)dk(o)。   
　　　　如图4，o1o1到p的第5可达距离为d(p,o1)d(p,o1)，o2o2到p的第5可达距离为d5(o2)d5(o2)。   
　　　　

**5) local reachability density：局部可达密度**   
　　　　点ｐ的局部可达密度表示为：

lrdk(p)=1/(∑o∈Nk(p)reach−distk(p,o)|Nk(p)|)lrdk(p)=1/(∑o∈Nk(p)reach−distk(p,o)|Nk(p)|)

　　　　表示点p的第k邻域内点到p的平均可达距离的倒数。   
　　　　注意，是p的邻域点Nk(p)Nk(p)到p的可达距离，不是p到Nk(p)Nk(p)的可达距离，一定要弄清楚关系。并且，如果有重复点，那么分母的可达距离之和有可能为0，则会导致lrd变为无限大，下面还会继续提到这一点。   
　　　　这个值的含义可以这样理解，首先这代表一个密度，密度越高，我们认为越可能属于同一簇，密度越低，越可能是离群点。如果p和周围邻域点是同一簇，那么可达距离越可能为较小的dk(o)dk(o)，导致可达距离之和较小，密度值较高；如果p和周围邻居点较远，那么可达距离可能都会取较大值d(p,o)d(p,o)，导致密度较小，越可能是离群点。

2.2 时序曲线预测

2.2.1 神经网络

在信息技术中，神经网络是在人类大脑中运行神经元后形成的硬件和/或软件系统， 这些技术的商业应用通常集中于解决复杂的信号处理或模式识别问题。自2000年以来，重要商业应用的例子包括支票处理的手写识别，语音到文本转录，石油勘探数据分析，天气预报和面部识别。

神经网络通常涉及大量处理器并行运行并按层排列的处理器。第一层接收原始输入信息 - 类似于人类视觉处理中的视神经。每个连续的层次都接收来自其前一层的输出，而不是来自原始输入 - 同样，距离视神经越远的神经元接收距离它较近的那些信号。最后一层产生系统的输出。每个处理节点都有其自己的小知识领域，包括它所看到的以及它最初为其自身编程或开发的任何规则。这些层是高度互连的，这意味着层n中的每个节点将连接到层n-1中的许多节点 - 它的输入 - 以及层n + 1中，为这些节点提供输入。输出层中可能有一个或多个节点，从中可以读取它产生的答案。

神经网络以适应性着称，这意味着他们在从最初的训练中学习时修改自己，随后的运行提供关于世界的更多信息。最基本的学习模型主要集中在对输入流进行加权，这就是每个节点如何权衡每个前辈输入的重要性。有助于获得正确答案的输入权重更高。

通常情况下，一个神经网络最初被训练，或喂食大量的数据。 培训包括提供输入并告知网络输出应该是什么。 例如，要建立一个网络来识别演员的面孔，最初的训练可能是一系列演员，非演员，面具，雕像，动物面孔等等的图片。 每个输入都伴有匹配的标识，如演员姓名，“不是演员”或“非人”信息。 提供答案可让模型调整其内部权重，以了解如何更好地完成工作。 例如，如果节点David，Dianne和Dakota告诉节点Ernie当前输入图像是布拉德皮特的图片，但是Durango节点说它是Betty White，并且训练计划证实它是Pitt，Ernie会减少它分配给的重量 杜兰戈的投入并增加了它给大卫，黛安娜和达科他的重量。

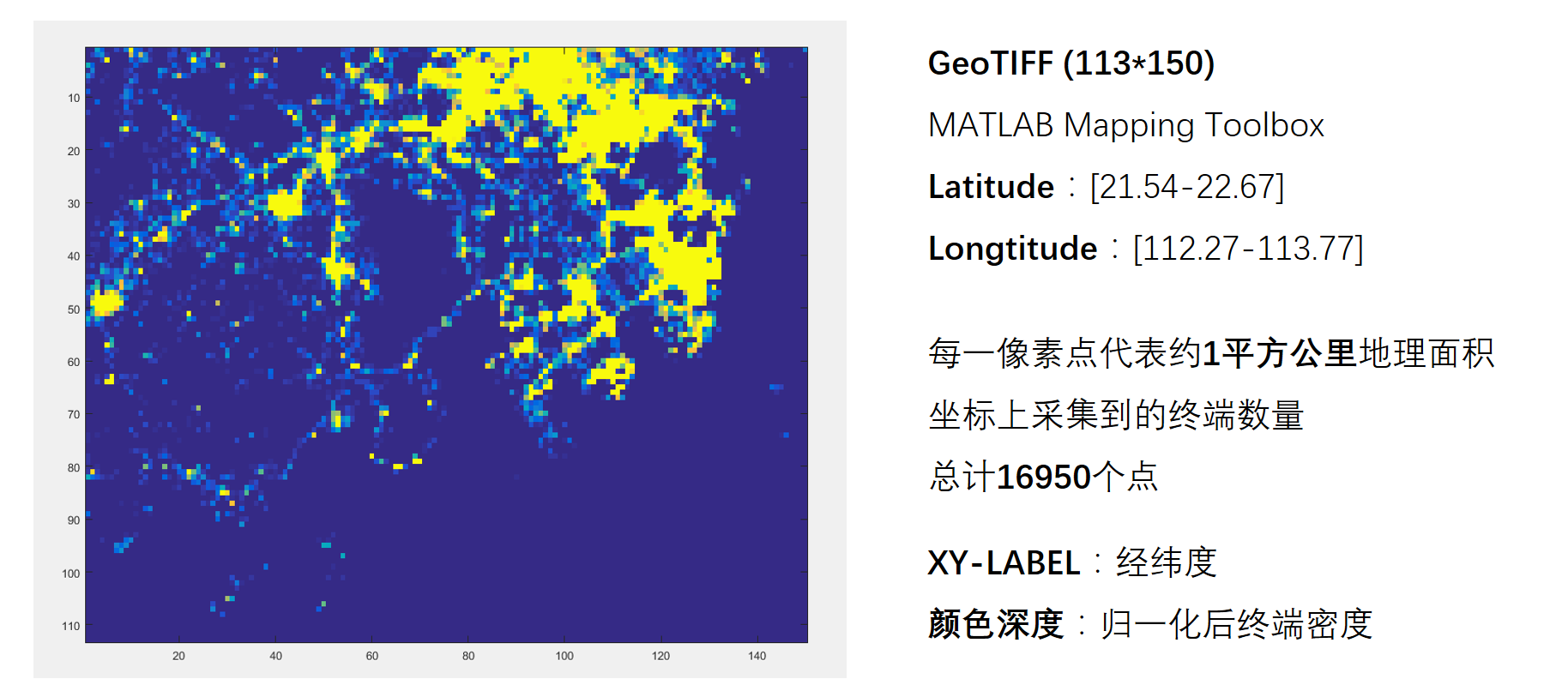
在定义规则和做决定 - 也就是说，每个节点根据自己从前一层的输入决定发送到下一层的内容 - 神经网络使用几个原则。这些包括基于梯度的训练，模糊逻辑，遗传算法和贝叶斯方法。他们可能被给予一些关于被建模空间中的对象关系的基本规则。例如，可能会指导一个面部识别系统，“眉毛在眼睛上方”或“胡须在鼻子下方，胡须在嘴上和/或嘴旁。”预加载规则可以使训练更快，并使模型更快更强大。但是它也建立在对问题空间性质的假设上，这可能被证明是无关紧要的，无益的或者不正确的，并且适得其反，这就决定了要制定非常重要的规则（如果有的话）。

1. 定位数据

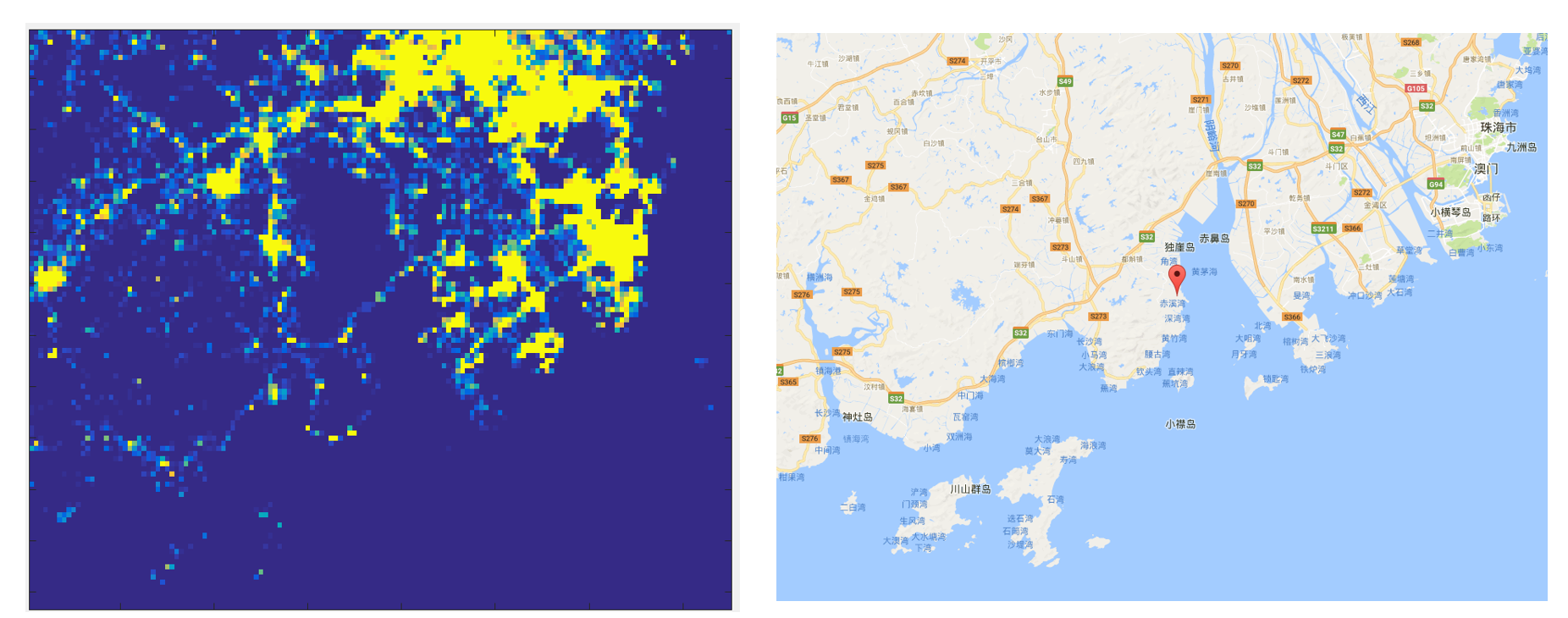
本章主要介绍了本课题所研究的腾讯定位数据的基本形式，并在MATLAB中作图分析其数据特征，包含时序特征及地理特征，最后根据分析结果讨论如何进行异常检测。

3.1 腾讯定位数据的形式

本课题所给定的腾讯定位数据是使用MATLAB的Mapping Toolbox生成的GeoTIFF格式的图像文件，每张图像文件的分辨率为113\*150，文件中的Reference信息包含图像所表示的地理位置信息。本类图像数据经过处理后可以确定该图像单位像素上的值表示实际地图上的0.01经度与纬度覆盖面积上（约为平方一千米）定位终端数量，其中横坐标表示经度，纵坐标表示纬度。根据上述地理信息可以在MATLAB中画出该区域的定位终端热力图，为表现特征，采用归一化后终端定位数据作出该区域定位终端密度热力图如下图所示**：**



将地理位置信息与实际世界地图进行比对，大致确定数据坐标为广东省珠海市沿海一带（如图）。同时，本数据集记录了8月14日至9月30日总计48天每天的每一小时区域定位终端数量。其中，该数据中已知的异常事件为8月23日的台风过境，由于台风袭来势必会导致地图上的终端定位数量发生显著改变，本课题通过分析该时段的终端定位数据来研究定位数据的异常检测方式。

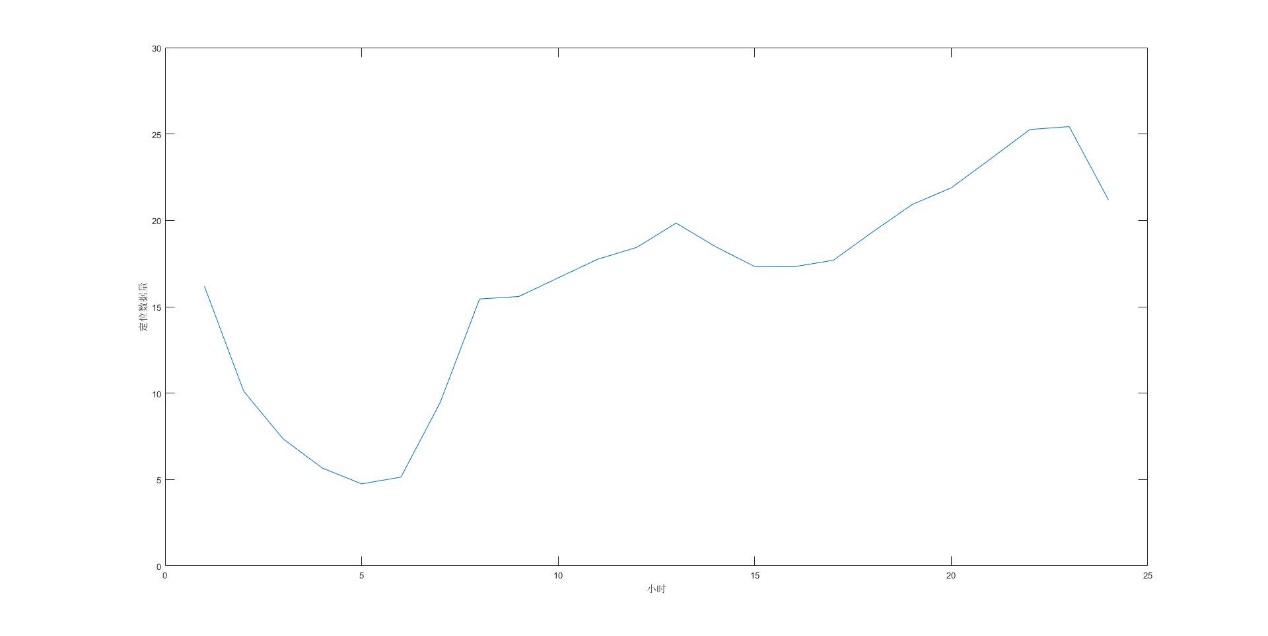


3.2 数据分析及应用

对于本课题，数据的维度涵盖时间与空间，直观上要分析单日的区域性台风影响带来的异常是困难的，我们首先需要对定位数据的规律进行分析，以便确定异常检测的算法思路。

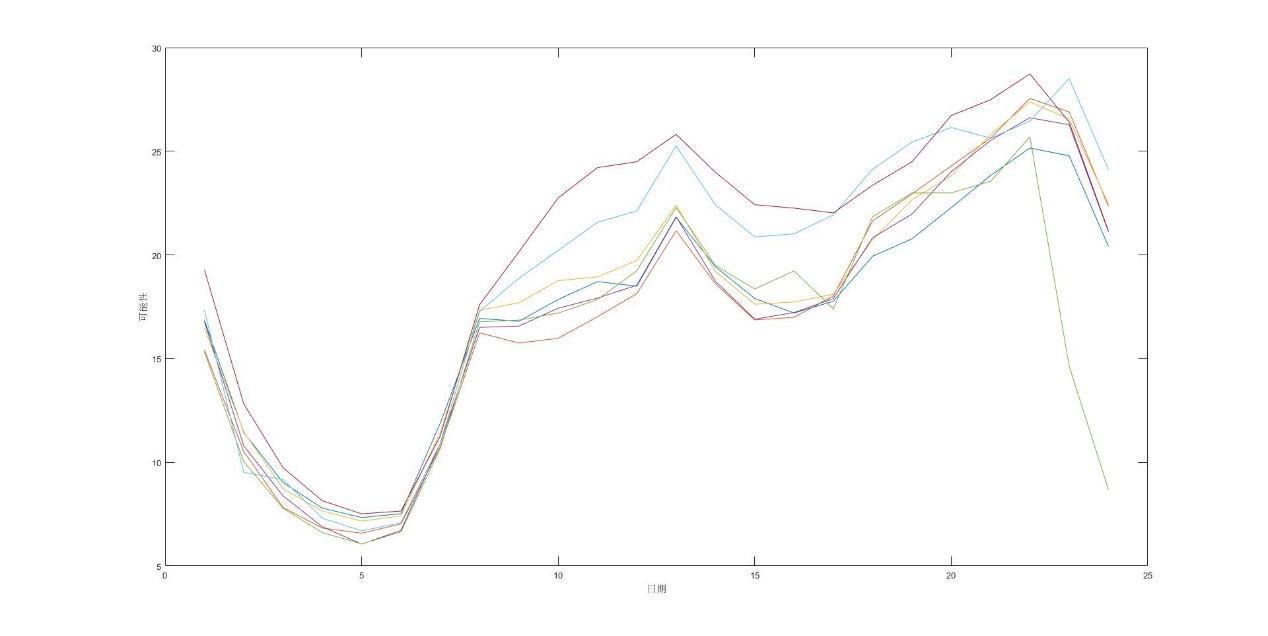
3.2.1数据的小时变化规律

定位数据在时间上以小时的单位进行采样，可以通过观察某地一天24小时的终端数量值得出定位数据的小时变化规律。为便于观察，应选取终端数量较多的区域从而得出普适规律，而在3.1中我们通过比较已经确定该区域的实际地理位置，可以选择图中人口密度相对较高的珠海市进行研究。将珠海市的地理坐标范围确定后，取该区域的定位终端数量平均值并绘制出其从午夜0点至次日午夜0点的小时变化曲线图，如下图所示**：**

****

由图中可以看出：午夜至清晨时间段为定位终端数量最少的时间段，大多数终端处于关闭状态，这是由于在这期间用户正处于睡眠状态造成的；而随着清晨至午间及晚间的推移，用户逐渐起床、工作、娱乐，定位终端数量也能观察到数量上的上升，而后从午夜开始再次下滑。

为确定此规律符合每一个正常的自然天，而不是工作日或休息日的特殊情况或是误采了某个节日的数据，再取该区域的定位终端数量平均值并在一张图内用不同颜色的曲线绘制出其一周每一天24小时内的小时变化曲线，**如图所示**

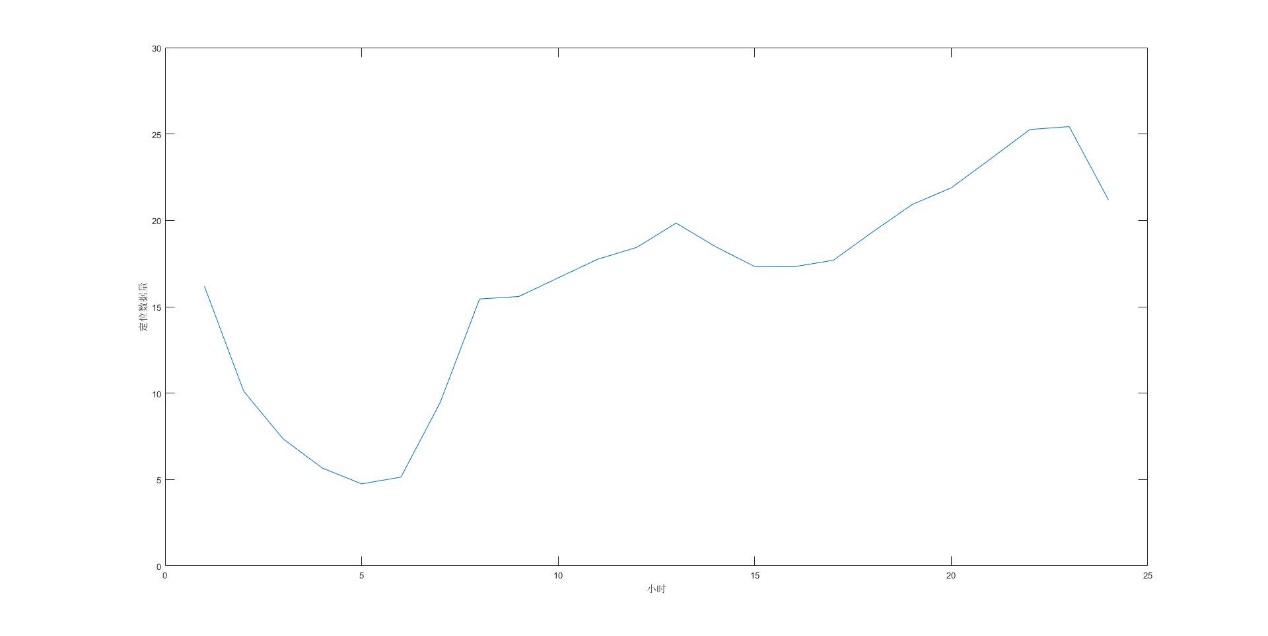
****

接下来我们再将一个正常天的24小时内终端数量变化的曲线与台风天的曲线进行比较，如下图所示。

由图中可以分析得到数据的小时变化规律：定位终端数据在确定的地点内一周中每天的变化趋势大致相同，而对于我们所已知的台风天异常，所造成的异常影响将不仅限于某小时带来的影响，而是会对整天各个小时的数据产生大的波动。

3.2.2 数据的日变化规律

在上一小节中，我们绘制出了台风天与正常天一天内24小时定位终端数量的变化并且观察得到台风天在一天内的值较正常值有较大的偏差。由于定位数据又在时空坐标上以自然天的单位进行采样，我们可以用同样的思路挖掘数据的日变化规律。为便于观察，同样选择图中人口密度相对较高的珠海市进行研究并选择一天当中定位终端数量较大的13:00时刻进行研究，取该区域每一天13:00的定位终端数量平均值并绘制出其在数据范围的48天内的日变化曲线图，如下图所示**：**

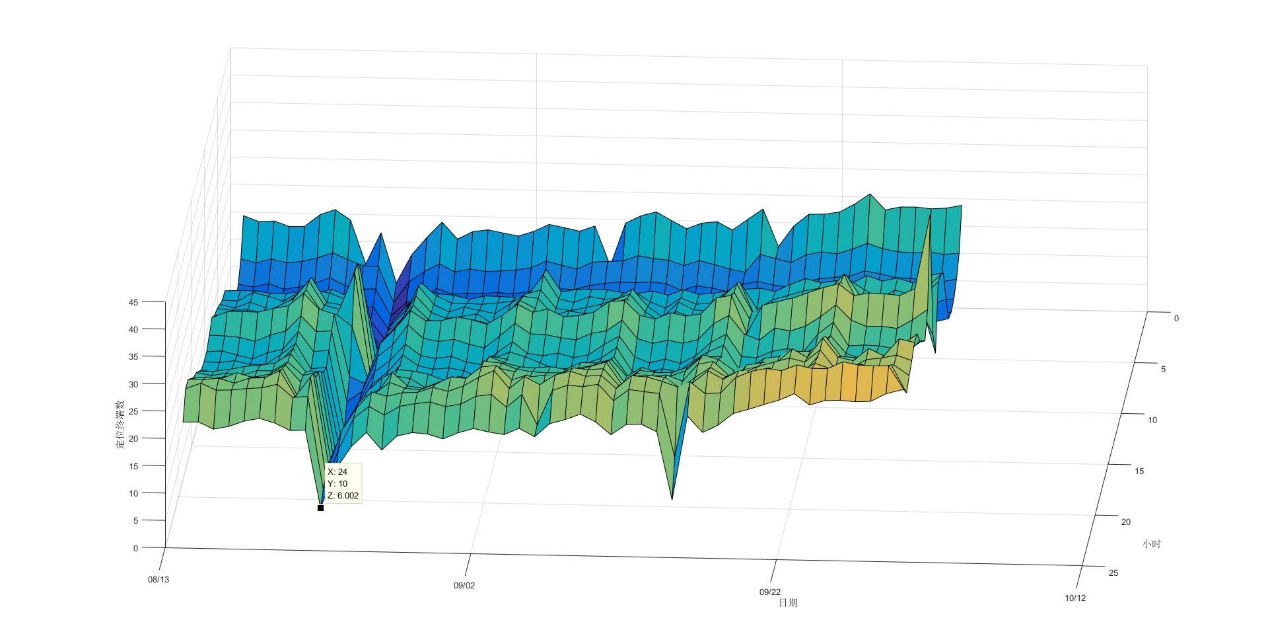
****

由图中可以分析得到：在该地的每天13:00时刻，台风天相较于正常天的定位终端数有明显的偏差，是上一小节中24小时的定位终端变化曲线中的13点时刻偏差值的天数扩展。

那么，如果能够通过选取任意的小时时间节点来代表整一天的定位终端数量值，台风天异常检测问题将会转化为每一天中的某确定小时的曲线异常检测问题。我们对此做进一步验证及讨论。

3.2.3 数据的总变化特征

将上述两章所分析的小时变化规律及日变化规律进行汇总，以X轴为数据范围内的自然日，Y轴为自然日内的每一小时，在Z轴绘出XY形成的＜日-时＞时间节点上的定位终端数量，如下图所示：

****

由该三维图的多角度观察可以分析得到：一天24小时内的定位数据变化确实大致相似，并且定位数据在8月23日中整体出现了一个明显的沟壑。但是，与前两小节的小部分时间节点分析不同的是，在这张图上某些小时时间点上出现了数据的突变暴露出来。如果仅依据每天中单个小时的数据对整天进行分析，过大的单小时点可能会对算法的判断产生误差；另外，由于图中仅显示了某固定点的定位数据图，需要对每个地点的时空特征进行统计，得到总体的判断依据。

3.3 数据预处理及异常分析策略

3.3.1 数据预处理

经过**3.2**对数据的分析，我们得到了定位数据的基本形式同时分析了其变化规律。在3.2中，我们选取的大多是极具代表性的区域（人口密度较高的珠海市）进行分析。但实际由于该定位坐标沿海，或是从任意时间节点上的定位终端矩阵或是绘制的区域热力图中也能观察得出：位于海面上的坐标终端数值存在大量接近零的点**，如图所示**。这些点无论对于分析数据规律或是检测异常都是冗余的，比如海面上某点两时刻的值从1到2有100%的变化，会极大地影响基于变化率的检测方法，需要将这些点进行剔除。



平均数是一个衡量区域内定位终端数量量级的基本方法，但是考虑到海面上可能会在某时刻突然出现了高额终端数这种极端异常情况，不能轻易地将其忽略。采用平均数阈值去衡量有效点可能会因为天数过多而将这种异常点舍去，因此本课题更适合采用最大值阈值的方法对数据进行预处理。

创建一个与定位数据地图相同大小的0-1矩阵表征定位数据图中像素点是否有效（以下称为有效矩阵），读取定位数据中每一个像素点在**所有时刻**的值，如果这些值中没有一个超过10（1平方公里的区域中没有一个时刻超过10个定位终端），则将有效矩阵相同位置处置为0，否则置1。经过这样处理后定位数据地图中只有约**2000多个点**有效，极大地加快检测速度的同时也避免了突变的错误舍去。

3.3.2 台风天异常分析策略

对于本课题所讨论的异常，应被归类为环境异常。环境属性即是时空坐标与地理坐标，行为属性是某地理坐标下在某时间点上的定位终端数量。本课题的检测目标即是输入所有的时空与地理坐标上的定位数据来检测出某一天的时空异常**。**

**3.2**中我们讨论了数据的时空特征，对于某固定的地理坐标，其可能会在某个点上出现。而在**3.3.1**中我们又对数据进行了预处理，减少了地理分析量。经过上述讨论，对于本课题所研究的台风天异常，对经过数据预处理后的筛选点进行时空维度上的曲线异常检测，使用每天的平均数据来分析，判断异常日期是哪一天或是全部为正常数据；再从地理上统计地图上所有已筛点的异常日期，如果地图上的大部分点都指向某一天存在异常的，即可认为该天是异常天。

1. 基于曲线分析的定位数据异常检测

经过在第3章中对数据的分析，我们将本身<时-空>的坐标分开分析，先对单个空间上的坐标点进行时空曲线异常检测，再统计空间上的规律，得出台风异常天的检测结果。在本章中，我们对每一种方法进行了设计与验证，查看其是否能够成功检出8月23日的台风天结果，并分析了各个算法的优势以及弊端。

4.1 基于小波的异常检测算法

4.1.1 离散序列小波变换

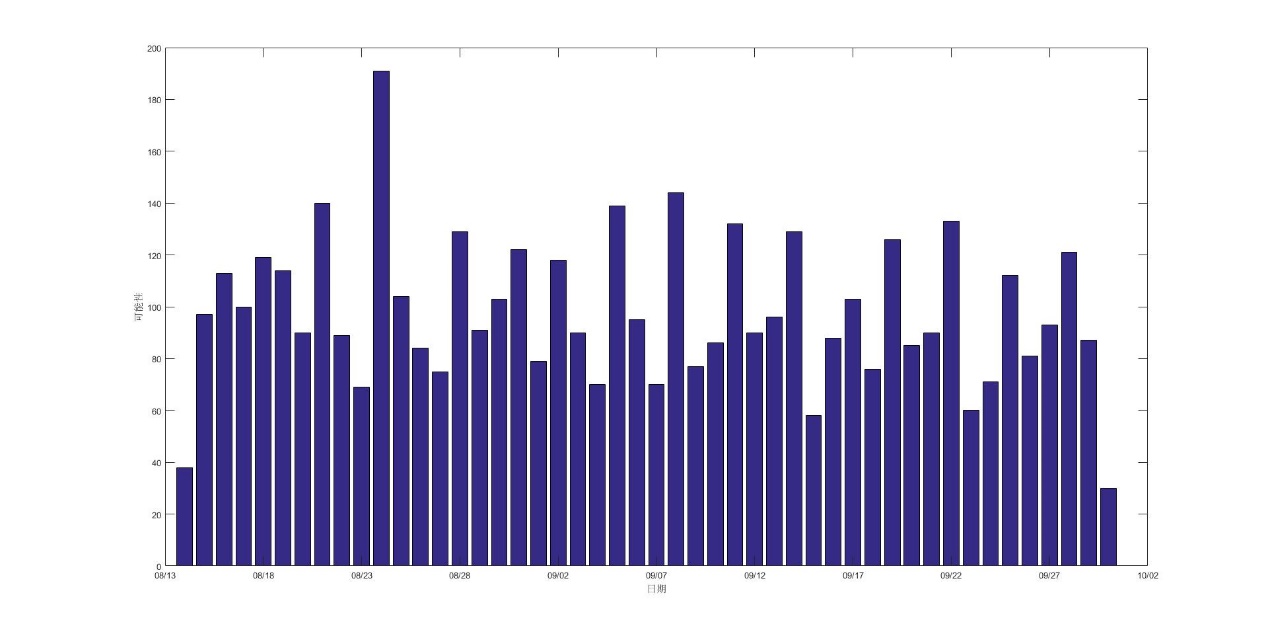
离散序列的小波变换就是基于著名的Mallat算法，离散序列值x与其第一层分解后的高频系数D1（细节部分Detail）的关系是x经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是x经过低通滤波器h滤波后再下采样；然后继续对低频系数A1进行第二层分解，依此类推，即离散信号x，经过多层分解后最后各分解系数合起来就是变换的结果。

对于本课题的数据，因为数据规模较小，直接采用一层分解即可从原始离散序列分离出有效高频部分D1，即原始信号的突变分量（异常分量）在D1中体现。对D1进行模糊处理并结合A1重建信号与原始信号差分，取最大值的横坐标（日期）即可得到该地理坐标下的异常日期。

4.1.2 结果分析

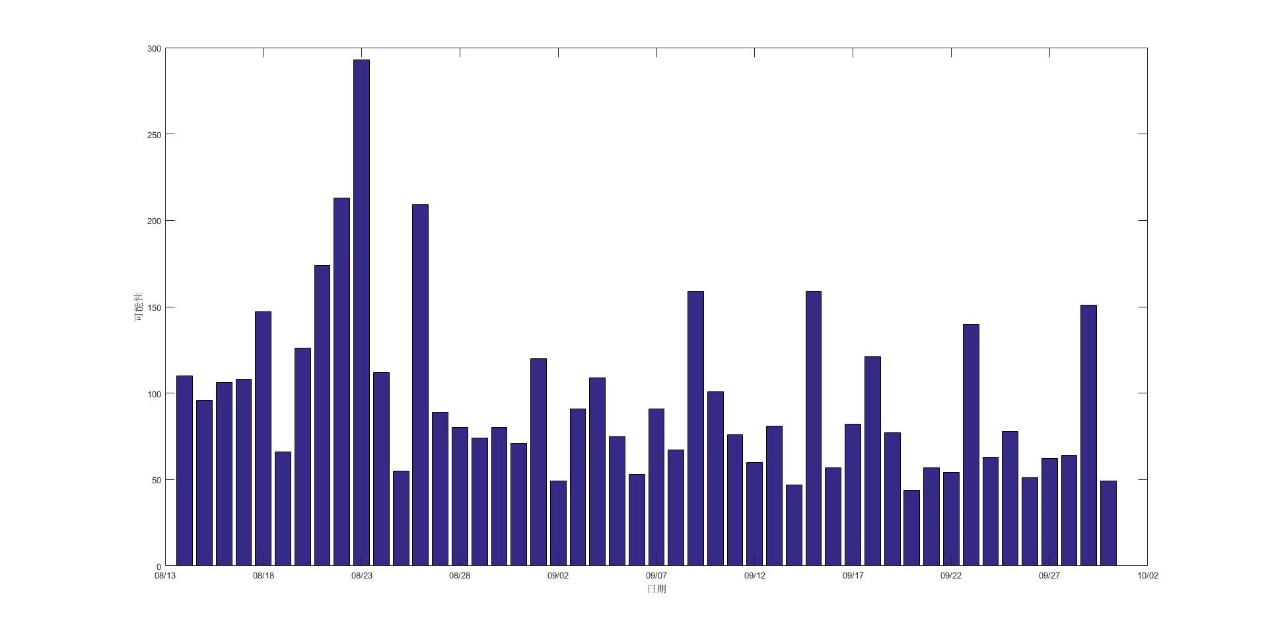
使用一层的离散小波变换成功检出了异常天（如图所示），但是有数量相对较大的误检测，并且小波分解计算量较大。

考虑到离散小波分解后是直接采用寻找最大值寻找异常天的横坐标的方法，在某些地理位置上其变化幅度较小，导致曲线本身就很平滑，采用小波变换后提取到的信号高频特征不明显，从而导致取最大值时发生错误造成了误检测。小波变换的确可以将离散的曲线信号中最为明显的部分突出，但后续处理的方法仍需改进。



4.1.3 改进算法

在4.1中我们讨论的小波变换能够将时序数据的高频分量提出，进而放大原始数据的噪声异常点。但由于在那一章中将异常放大后直接采取取最大值的策略导致很多本身没有异常的点也被错误的认为是异常点。所以，我们对该方法进行了优化，将原始时序数据经过小波分解后再进行差分运算，得到异常天的日期。同时，我们又对4.4中提到的原始差分运算进行改进，从相邻点的差分拓展到对相邻三点间的差分，减少异常点对周围点差分的影响。



4.2 基于极大似然估计的异常检测算法

4.2.1 极大似然估计与3σ准则

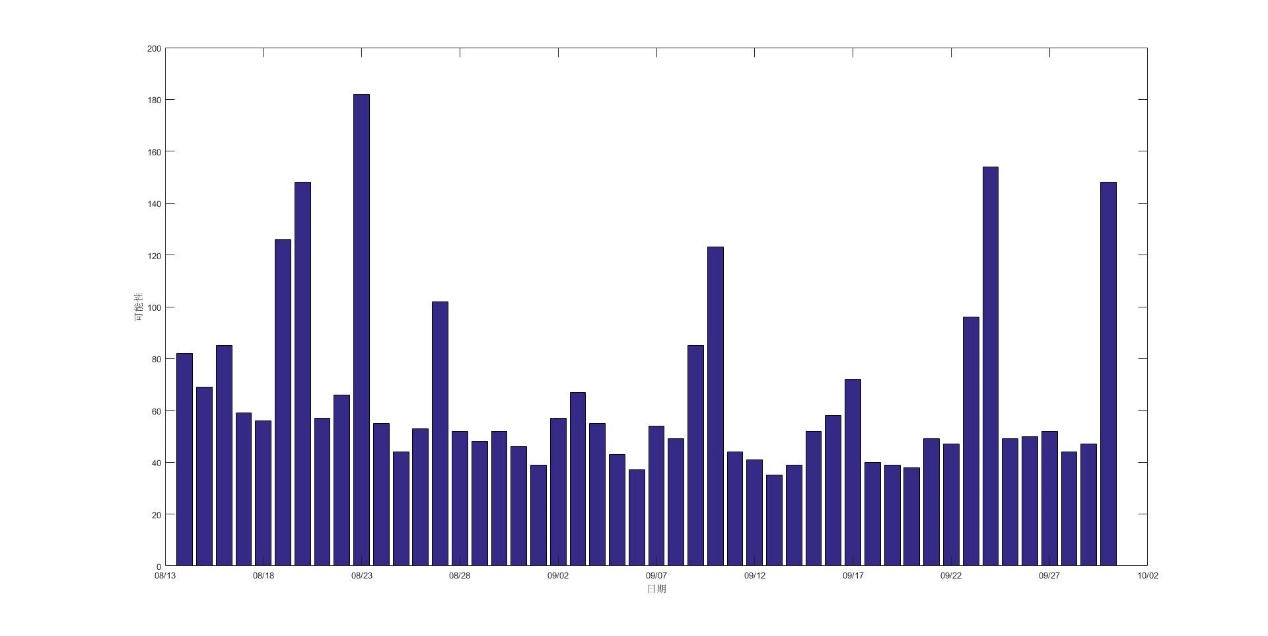
在第2章中我们讨论的极大似然估计的基本含义，其是用来估计一个概率模型的参数的一种方法，即通过若干次试验，观察其结果，利用试验结果反推最有可能（最大概率）导致这样结果的参数值。经过第3章的讨论，我们已知某地理坐标上的定位终端数在没有异常事件到来的情况下浮动规模应大致符合正态分布。基于此假设后，对于某地的时间序列求解最大似然估计，求解正态分布下似然方程得到唯一解（μ\*, σ\*2）。

在有了似然估计的解之后，我们得到了数据本身的一种拟合分布情况。在正态分布下，数值分布在（μ-3σ,μ+3σ)中的概率为0.9973。可以认为，正态分布中Y（在本课题中为定位终端数量）的取值几乎全部集中在（μ-3σ,μ+3σ)区间内，超出这个范围的可能性仅占不到0.3%。本课题中所涉及的异常数据如果较正常数据偏差大，很有可能会落在此小区间中被检出，即可得到异常天的日期，否则认为不存在异常天。

4.2.2 结果分析

使用极大似然估计拟合数据的分布，并基于3σ准则将异常值筛选出后，得到了如图所示的异常天分布情况。本算法仍然检出了异常天，但误检率仍然较高，且由于要求解每一个坐标点的似然估计方程，计算量相对于小波变换更加大。

考虑到极大似然估计是基于现有的数据对原始分布进行拟合，需要较为庞大的数据量支撑以便充分拟合，才能忽略个别噪声的影。而在本课题所涉及的数据中，数据量较小且这些用来估计的现有数据中也包括了异常的数据，如果异常值偏离很大，估计出的参数会极为不准确，从而导致误检。另外，为了便于计算，本算法认为原始数据基于正态分布，实际情况下需要进行长时间的统计，对区域的定位终端数量有一个充分的采样继而判断该地的数据符合哪一种分布，最后对这种分布的参数进行极大似然估计，才能得到较为准确的模型以及判断异常的条件。

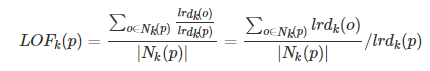


4.3 局部异常因子检测算法

4.3.1 局部异常密度

点ｐ的局部可达密度表示为：

  
　　　　表示点p的第k邻域内点到p的平均可达距离的倒数。   
　　　　注意，是p的邻域点Nk(p)Nk(p)到p的可达距离，不是p到Nk(p)Nk(p)的可达距离，一定要弄清楚关系。并且，如果有重复点，那么分母的可达距离之和有可能为0，则会导致lrd变为无限大，下面还会继续提到这一点。   
　　　　这个值的含义可以这样理解，首先这代表一个密度，密度越高，我们认为越可能属于同一簇，密度越低，越可能是离群点。如果p和周围邻域点是同一簇，那么可达距离越可能为较小的dk(o)dk(o)，导致可达距离之和较小，密度值较高；如果p和周围邻居点较远，那么可达距离可能都会取较大值d(p,o)d(p,o)，导致密度较小，越可能是离群点。   
　　**6) local outlier factor：局部离群因子**   
　　　　点p的局部离群因子表示为：

  
　　　　表示点p的邻域点Nk(p)Nk(p)的局部可达密度与点p的局部可达密度之比的平均数。   
　　　　如果这个比值越接近1，说明p的其邻域点密度差不多，p可能和邻域同属一簇；如果这个比值越小于1，说明p的密度高于其邻域点密度，p为密集点；如果这个比值越大于1，说明p的密度小于其邻域点密度，p越可能是异常点。   
　　现在概念定义已经介绍完了，现在再回过头来看一下lof的思想，主要是通过比较每个点p和其邻域点的密度来判断该点是否为异常点，如果点p的密度越低，越可能被认定是异常点。至于密度，是通过点之间的距离来计算的，点之间距离越远，密度越低，距离越近，密度越高，完全符合我们的理解。而且，因为lof对密度的是通过点的第k邻域来计算，而不是全局计算，因此得名为“局部”异常因子，这样，对于图1的两种数据集C1和C2，lof完全可以正确处理，而不会因为数据密度分散情况不同而错误的将正常点判定为异常点。

4.3.2 结果分析

采用了

4.4 基于差分的异常检测算法

4.4.1 差分算法

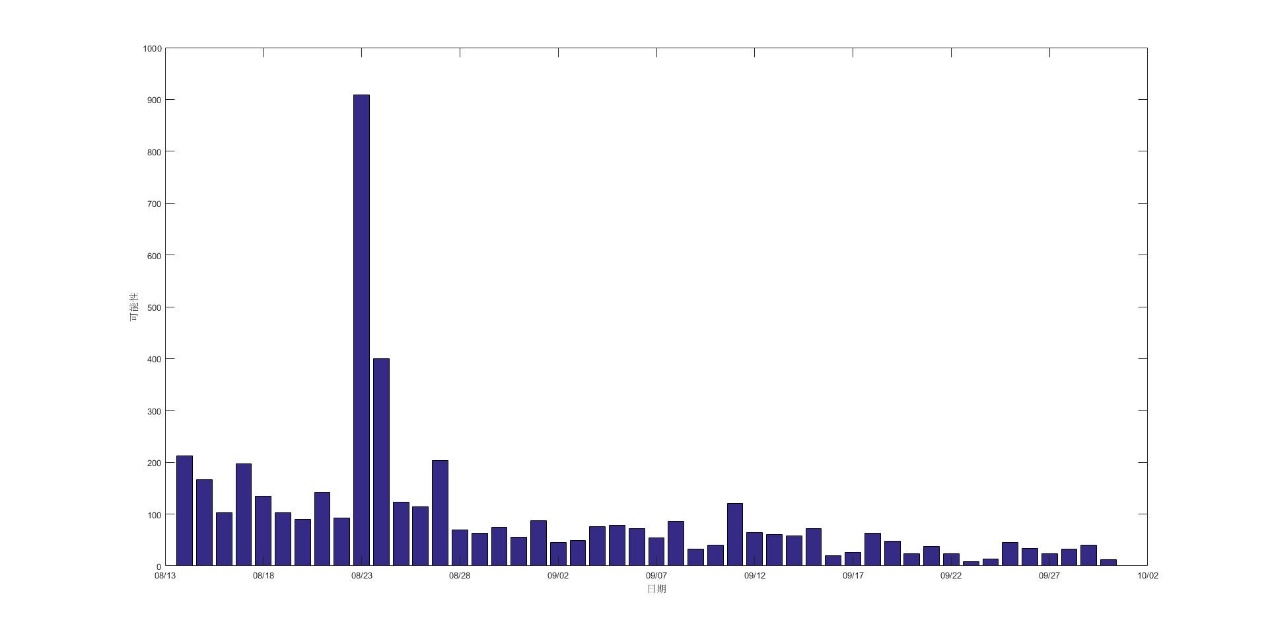
异常的定义是某个数据严重偏离正常数据的范围之内，那么同样也可以认为在时序数据中，异常数据较其周围的数据有很大的波动。使用差分算法计算每一个点与其左右两个时间节点点上的数据浮动比例，当这个比例超过某种阈值后，即可认为该点是异常的。

4.4.2 结果分析

采用20%的变化率阈值的差分算法得到了如图所示的异常天分布情况。本算法成功检出了异常天误检率已经较低，并且由于本算法完全是线性运算，计算量较低。

差分算法是一种更加针对平稳的时序数据算法，因为它只考虑每一个数据点相邻时序上的点，所以对于时序上突然抖动的数据更加敏感。而且，使用变化率来衡量的差分算法也对小数据的微量浮动很敏感。在本课题中，因为已经事先对数据进行预处理，排除掉了小数据的干扰才使得差分算法准确度较高。在实际情况中，更应该使用变化率加上变化阈值的双条件来判定数据是否异常。

另外，本算法中所涉及的差分采取的是时序上左右两点间的差分值，真实异常点的附近几点也会受到异常点的影响导致可能被归类为异常点，应扩大差分的范围，减少异常点对周围点的影响。并且，对于实时的异常判断来说，不存在时序右侧的差分，可以将其替换为时序前一点与其前一点之间的差分，再求它们之间的平均值。



4.6 曲线异常检测算法总结

这一段还在优化中

1. 定位数据的预测分析

本章主要基于第4章中的异常检测基础，提出了时序数据的预测估计方法，并根据神经网络训练模型预测当前时间节点的值与实际值进行比对，进一步讨论了异常检测的另一种思路

5.1 异常的检测与预测

时间序列是根据时间顺序得到跟时间相关的变量或者参数的观测数据[1]。对时间序列的研究主要是挖掘其中有价值的信息，找到其中变化的内在规律[2]。在第4章中我们所讨论的问题都是基于现有的<空-时>数据，从其中检出哪一天存在区域性的异常，检测是一个判别的过程，是从已知数据中检测出异常的模型，这种检测方法可以用于数据清洗。同时，异常检测在生活中也多用于即时的判断，例如网络流量的异常检测可以对DDoS攻击进行相应，使网络运营商做好应急预案。对于本课题所研究的定位数据来说，如需实际应用那么检测方法更加强调实时性，即输入一组新的数据后基于现有的数据能够判别新数据是否存在异常。

时间序列预测是指根据现有的和历史的时间序列的数据，建立能反映时间序列中所包含的动态依存关系的数学模型[3]。我们可以通过建立模型来预测新的数据理应符合什么区间，并与实际数据进行比较，判断其是否异常。

5.2 基于动态神经网络的定位数据预测

5.2.1 动态神经网络

神经网络是一门重要的机器学习技术，它是一种模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，主要用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。

大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统。通过训练样本的校正，对各个层的权重进行校正（learning）而创建模型的过程，称为自动学习过程（training algorithm）。具体的学习方法则因网络结构和模型不同而不同，常用反向传播算法(Backpropagation/倒传递/逆传播，以output利用一次微分Delta rule来修正weight)来验证。通过这样学习（训练与验证）的过程，它可以对目标函数进行相对完整的模拟。

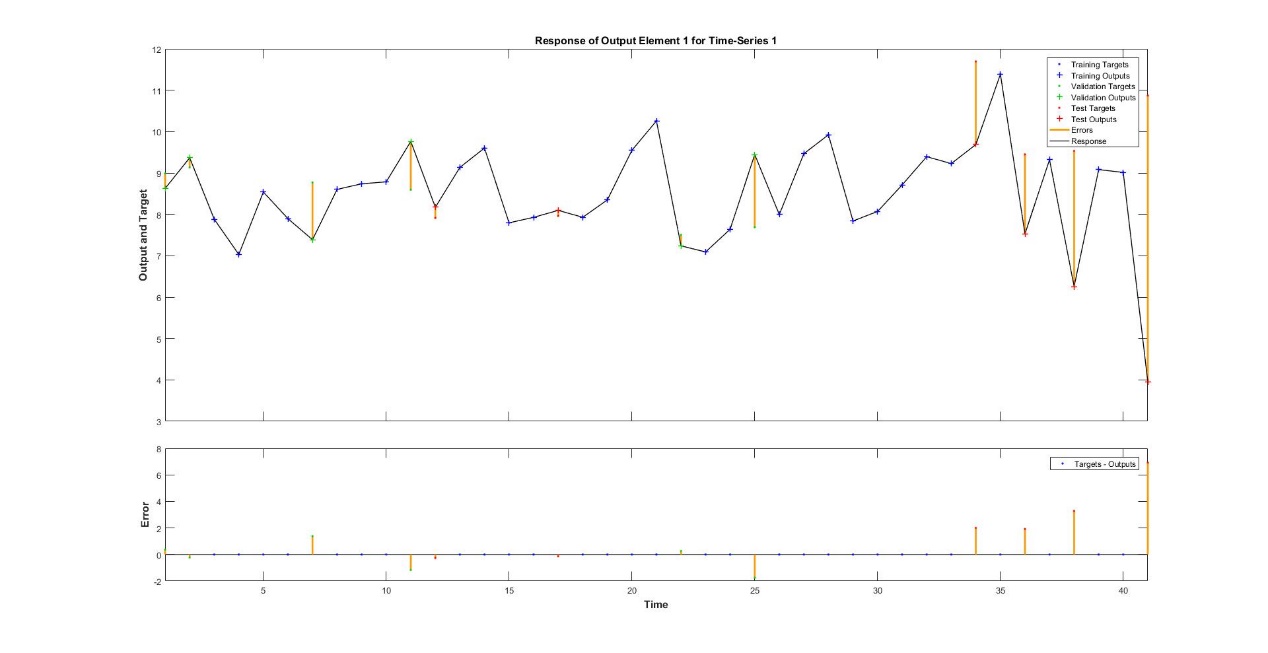
神经网络又按照是否含有延迟或反馈环节可以分为静态神经网络和动态神经网络，含有延迟或者反馈环节的神经网络被称为动态神经网络。静态神经网络由前向传播的神经元组成，连接仅馈送到后续层,因为没有回馈的因素（即同一列神经元间没有相连的关系），其主要被用于静态判断或预测，例如卷积神经网络被用来识别静态图像。动态神经网络中一个显著的特点就是神经元间的输出回馈到前序神经元中，神经元之间的联系使得动态神经网络又可以解决静态神经网络无法解决的时序问题，因为 RNN 包含循环，所以它们可以在处理新输入的同时存储信息。这种记忆使它们非常适合处理必须考虑事先输入的任务（比如时序数据）。在4.4中，我们曾用差分算法检测异常值，差分算法的本质在于通过相邻时间节点的浮动变化率对异常点进行判断，即根据变化趋势来检测是否存在异常。上述动态神经网络也需要使用训练出的模型并根据前几项时间节点上的值，来预测当前时间节点上的值。

对于本课题所讨论的定位数据异常检测，动态神经网络起到了先预测当前值的作用，而对于这个预测值，我们只需要和当前实际值进行比较，如果他们相差的幅度达到一定规模，则认为当前值存在异常，应当启动应急方案来应对。

5.2.2 定位数据预测

本章节将使用MATLAB软件中的神经网络工具箱，以某地区的过往定位数据时间序列为输入，该地区当前时间节点的定位为输出，依据输入的过往时空数据构建神经网络模型，以时间序列预测方法来进行当前定位数据量预测。借助神经网络的非线性问题处理能力，根据不同的实际情况，对数据进行训练，并与实际定位数据对比，验证方法的有效性。

在4.2的极大似然估计中，我们提到异常的数据会对极大似然估计造成很大的误差，本章我们所讨论的定位数据预测同样需要考虑这个问题，应当将异常的数据进行剔除，使用正常的数据训练网络结构并使用它来预测当前时间节点上的值。接下来的部分和第4章一样，将经过预处理后的定位数据导入MATLAB中，使用ntstool命令进入时间序列工具箱进行训练。采用动态神经网络有反馈模型，网络训练时把数据分为三类：训练数据、验证数据和测试数据，三者比例设置为：70%、15%、15%。由于每一个自然周的变化大致相似，所以采用了前7个点作为预测的前序值。通过训练数据和验证数据来训练神经网络的模型并自动反向传播调整网络参数。



**分析结果仍在补充中**

1. 基于区域的定位数据异常检测

前述几章我们主要讨论了基于腾讯时序定位数据的异常检测及预测算法，可以对异常发生的整体区域进行分析，在第3章中我们也提及该定位数据中存在大量海面上的无效数据，实际上发生异常的可能只是部分区域。在这一章中，我们根据已知的异常时间点，和正常时间点的整体区域，去寻找具体发生异常的子区域。

6.1 基于图像的异常区域检测

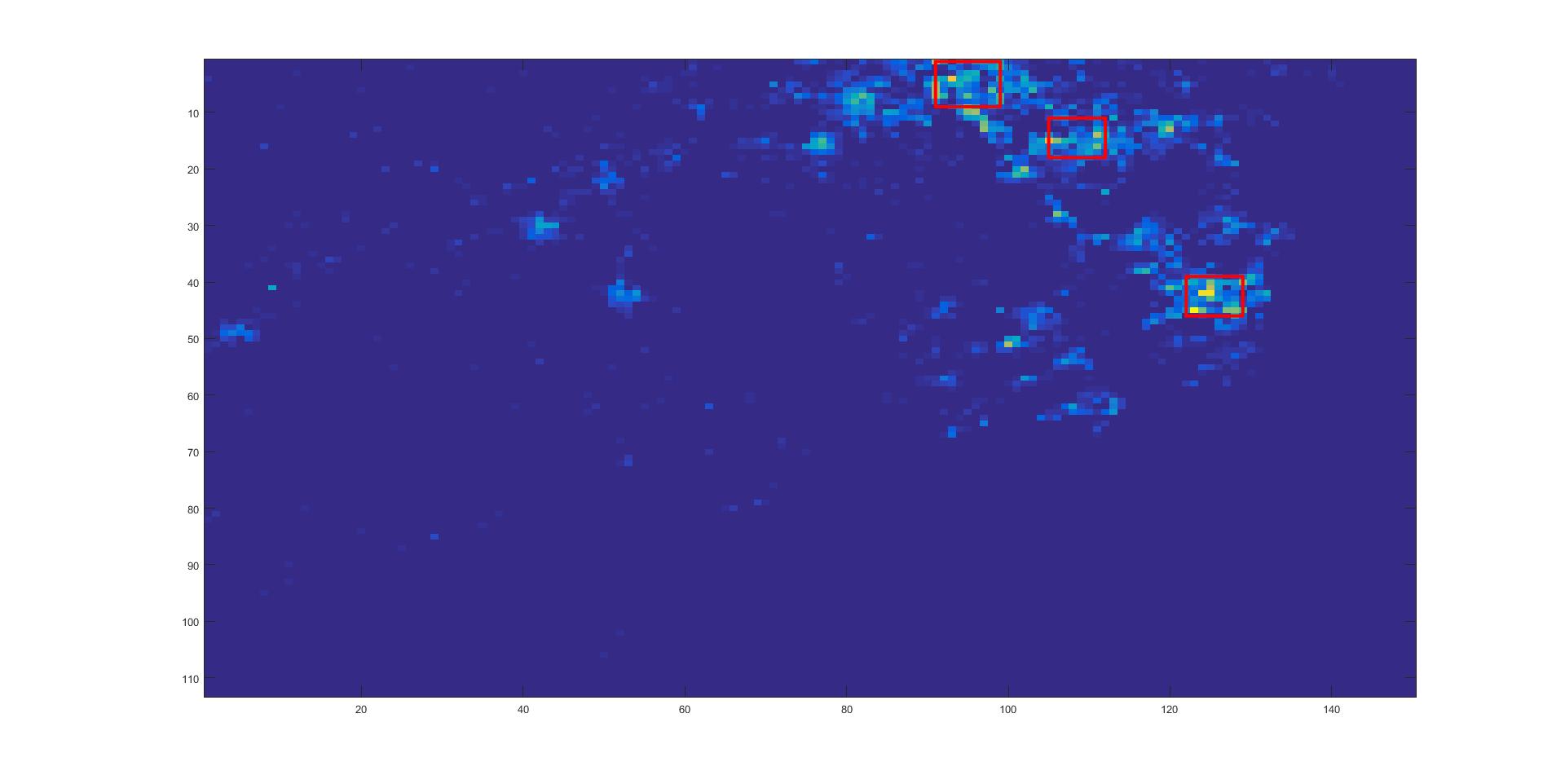
6.1.1 相邻帧间差分法

相邻帧间差分法是一种视频分析上检测物体运动的方法。由于摄像机采集的视频序列具有连续性的特点。如果场景内没有运动目标，则连续帧的变化很微弱，如果存在运动目标，则连续的帧和帧之间会有明显地变化。帧间差分法(Temporal Difference)就是借鉴了上述思想。由于场景中的目标在运动，目标的影像在不同图像帧中的位置不同。该类算法对时间上连续的两帧或三帧图像进行差分运算，不同帧对应的像素点相减，判断灰度差的绝对值，当绝对值超过一定阈值时，即可判断为运动目标，从而实现目标的检测功能。

6.1.2 高浮动区域检测

对于本课题，帧间差分法所实现的功能即是检测出异常导致的子区域定位数据的剧烈抖动。由于台风过境这类异常事件的产生，使得图中某一部分的定位数据会产生较大的变化，这些变化可以使定位数据产生的图像之间差分后在图中明显观察得到。

我们已知并且在第4章中检测出8月23日定位数据范围内存在异常，我们使用滑动窗的方法去检测高浮动区域，使用可自动伸缩的正方形滑动窗对整个范围内进行搜索，标记出几个密度最高的区域，在图中显示如下：



1. 总结与展望

7.1 内容总结

现代科学的发展，使得传感器网络越来越普及，让大数据分析成为可能。在这些数据之中，有一种数据尤为重要，其存在着很明显的时序性，描述着随着时间的推进中，某种事物的变化规律，这种数据常被称为时序数据。时序数据分析是对时间序列进行系统的分析并且简历合理的模型。其主要目的是考虑数据动态的波动情况并预报未来发生的事件。本文所讨论的定位数据就是基于某固定的空间坐标上的时序数据点，通过这种定位数据的时序分析，可以归纳总结出该地的人流变化特征。而对于存在明显规律的时序数据，我们可以对其进行建模并预测当前或未来时间节点上的值。

但是，时序数据中经常会出现这样一些的观测点，它们的数据值较同性质时段上的数据存在着明显的偏差，我们称其为异常点。异常点的分析以及处理十分有意义：一方面异常点对于时间序列的建模来说是干扰很严重的噪音，较大的异常会使得拟合模型造成极大的偏差，需要将这些异常点检出并且删去使得模型的训练及预测更加精准；而另一方面，时序定位数据中的异常又反应出了该时间点上一定存在着某种外界干扰，比如定位数据急剧降低可能是由于台风过境人群大量躲在遮蔽物处而造成的。对于台风这种自然灾害，如果能根据时序定位数据的预测以及异常的判断，则能很好的做好预警，显得尤为重要。本文根据以上思路，基于腾讯的定位数据，对时序异常检测及相关内容完成了以下工作：

1. 数据分析：对给定的腾讯定位数据有了基本的背景认识以及为了后续的分析方便进行了预处理。给定的腾讯定位数据大致地理坐标为广东省珠海市沿海一带，而明确探查出的异常是某天的台风过境；由于地图上涵盖一大部分海面，而海面上的定位数据（定位终端数量）几乎为零，分析这些数据毫无意义，故通过最大值判断的方法进行了舍去；最后，通过数据的可视化，我们得出了通过分析每天确定的某一小时来反应整天的定位数据信息，并统计分析图中所有地理坐标上的定位数据信息来确定某天是否存在异常的结论。

2. 曲线异常检测：首先我们需要能够检测出异常，而从数据的分析中我们将复杂的时序定位数据最后压缩成一条曲线来做异常判断；其次，经过基于统计、密度、差分、频域的几种曲线分析策略设计异常检测算法，我们都将时序曲线中的异常点找出；最后，基于上述几种方法，我们分析了其在时序定位数据中的适应性并对于时序预测及其他细节的问题进行了讨论。

3. 时序数据预测：首先，我们简单介绍了神经网络，并且强调了其对于时序数据的建模有着很大的帮助，它能够经过数据的训练来对网络结构中的参数进行调整，从而使网络近似趋近于时间序列的实际规律；其次，我们对本课题所研究的腾讯时序定位数据中一部分进行了训练，并调整了网络参数使得其在另一部分数据上的预测结果更加精准；最后，我们使用该预测模型对台风过境的定位数据进行了预测及判断，成功检测出台风临近时人流密度的大幅波动。

4. 区域的定位数据异常检测：首先，前述几章的异常检测方法反应出了整个定位数据区域内的异常情况，在本章中基于上述检测结果，我们将研究重点放在检测区域中哪一块小区域出现了大浮动导致了异常；其次，提出了图像相邻帧间差分法和聚类相结合的方法，寻找异常天中异常变化幅度最大的部分，用矩形窗圈出；最后，根据实际情况对上述结果进行了分析。

7.2 未来展望

本文的实验表明，在基于腾讯地图的时序定位数据上，我们能够成功检出已知的台风天异常并建立预测模型达到实时异常检测的目的，并且能够根据异常的时间点找出异常子区域。但对于实际定位数据的分析来说，本文在以下方面可以进行改善：

1. 高地理精度定位数据的异常检测：本文所研究的课题所涉及的定位数据是大区域的粗精度数据，一个像素点约代表平方一千米以上的量级，算法对于分析大范围内的整体异常是有效的。但实际情况中有时研究者所接触到的定位数据是更高精度的，此时本文所讨论的对于整体区域的异常检测不再有效，高精度下应该更关注地理位置上的部分区域，应在本文的第六章基础上做更进一步探索。另外，本文对数据所做的简化处理也无法再适用于高精度的数据，无论是时空上或是地理上高精度的定位数据有更多信息可以挖掘，无法直接从统计角度进行整体分析，应当对数据进行更加细致的分析研究可实现的算法。

2. 大规模数据的曲线异常检测：本文所实现的曲线异常检测算法中，对于本课题所研究的小规模数据量效果甚微，其中部分算法，例如极大似然估计法，当数据量足够大时可能会表现地足够好。同样，对于第5章所讨论的时序数据预测，数据量小时很可能导致网络模型过拟合，对于数据的拓展预测不利。

[1]田铮.时间序列的理论与方法.北京:高等教育出版社，2001.

[2]E. KEOGH.Data mining and machine learning in time series database[C]//Proc of the 5th Industrial Conference on Data Mining(ICDM), Leipzig. 2005.

[3]刘大同.基于Online SVR的在线时间序列预测方法及其应用研究[D].哈尔滨工业大学.2010: 1~3.