一种精细化模　　式方法，用于从移动设备位置数据中识别数据振荡

Qianqian Sun (qsun12@umd.edu**)**
研究生助理，土木与环境工程系，马里兰大学，美国马里兰州大学公园市格伦·马丁楼1173室，邮编20742。开放研究者与贡献者身份识别码：
0000-0003-3684-4603

Aref Darzi (adarzi@umd.edu**)**
教师助理，土木与环境工程系，马里兰大学，美国马里兰州大学公园市格伦·马丁楼1173室，邮编20742

Yixuan Pan（ypan1003@umd.edu）通讯作者研究科学家 土木与环境工程系 马里兰大学，格伦·马丁楼1173室，大学公园市，马里兰州20742，美国

摘要 近年来

，被动采集的GPS数据已广泛应用于各种交通研究，例如公路运行状况监测、出行行为分析和出行需求估算。尽管其具有多项优势，但其中一个问题是数据振荡（又称异常值或数据跳变）。这一问题不容忽视，因为它们可能扭曲移动模式，并导致错误或有偏差的结论。对于由GPS数据驱动的交通研究而言，通过去除数据振荡造成的噪声来确保数据质量无疑十分重要。大多数基于GPS的研究仅通过检查高速情况来去除振荡。然而，这种方法可能会将正常点误判为振荡。一些其他研究专门讨论GPS数据中异常值的去除，但它们都有局限性，并不适用于被动采集的GPS数据。许多研究已针对蜂窝数据或蜂窝基站数据中的乒乓现象进行了充分开发，但被动采集的GPS数据中的振荡在模式多样性和复杂性以及不确定性方面都大不相同。现有方法不足以处理被动采集的GPS数据，也不适用于此类数据。本文旨在处理被动采集的GPS数据中的振荡点。本文通过识别振荡所呈现的异常移动模式，提出了一组启发式方法。所提出的启发式方法很好地适应了被动采集的GPS数据的特征，并且能够适用于不同规模的研究；与现有方法相比，其计算成本也较低。

关键词：数据振荡、数据跳变、异常值、被动采集的GPS数据

1. 引言大型流式 GPS

数据（格式为设备标识符、时间戳、纬度和经度）正日益应用于许多领域。尤其是在交通运输领域，GPS数据作为现实世界出行的表征，已展现出其在支持出行相关研究方面的价值，例如公路性能评估、出行统计估算和出行需求估算。在传统调查因耗时或成本过高而无法开展或不可行的情况下，GPS数据可作为替代数据源。对于由GPS数据驱动的研究而言，仔细进行数据准备至关重要，这是确保出行推断可靠的关键。否则，推断可能会受到削弱或产生偏差 [1, 2, 3]。GPS数据准备中的一项重要工作是检测并移除数据振荡，即异常值或数据跳变。尽管数据振荡（与异常值一词交替使用）主要是在蜂窝数据或蜂窝基站数据的语境下讨论的，但GPS数据中也存在这种现象。然而，目前用于检测GPS数据中异常值的方法要么不可靠，要么关注异常轨迹而非异常点；而为蜂窝数据开发的方法也不适用于GPS数据。本文旨在识别被动采集的GPS数据中的数据振荡，并以点而非轨迹作为基本单位。本文特别讨论被动采集的GPS数据，因为这类数据在实践中应用更为广泛。被动采集的GPS数据主要有两个来源：直接通过全球定位系统（GPS）采集，以及间接通过基于位置的服务（LBS）采集。由于LBS数据依赖用户或平台触发定位服务，被动采集的GPS数据在不同时间窗口或不同位置可能呈现不同的数据特征（例如频率、空间密度）。这种多样性增加了区分振荡与正常点的难度。大多数异常值检测方法都是基于单个变量的统计分布（例如正态分布、伽马分布）进行统计检验 [4, 5, 6]。这种方法不适用于被动采集的GPS数据，因为这类数据通常不遵循特定分布。目前，轨迹数据或GPS数据的异常值检测方法可分为基于速度的方法、基于距离的方法、分区检测法、基于密度的方法、基于聚类或分类的方法，以及基于模式的方法。

基于速度的方法只是移除与前一个点相比速度极高的 GPS点。研究中采用了几种不同的速度阈值，例如 200 km/hr [7] 和 200 mph [8]。这类方法并不可靠。实际上，被动采集的GPS数据中的数据振荡情况要复杂得多。仅凭一个变量——速度——不足以识别振荡，甚至可能将正常点误判为振荡。例如，某点相对于前一个点的速度极高，但该点实际上可能是正常点，而前一个点才是真正的数据振荡。该方法可能会直接移除后一个点，甚至继续移除其后几个点。这种方法可能造成许多假阳性案例。之后，有人提出了一种改进的基于速度的方法 [9] 。该方法仍以不合理的高速（155 mph）为依据，作者提出使用递归前瞻滤波器和卡尔曼滤波器来移除异常值。递归前瞻滤波器首先体现了递归检查的价值；其次，当一对点显示出异常高的中间速度时，通过检查这两个点与第三个后续点之间的距离，判断其中哪一个是异常值。这两种思路为本文提供了参考。卡尔曼滤波器假设 GPS点是来自线性动力系统的车辆轨迹，其应用范围有限，因为实际中的GPS数据并非仅与车辆有关。本文旨在提出一种不局限于特定出行方式的通用方法。

基于距离的异常值检测方法 [6] 是早期一项从轨迹数据中检测异常值的研究。该方法首先用包含一组移动性特征（即起点和终点、点数、航向和速度）的多维向量来描述每条轨迹。然后，通过对两条轨迹向量之间的差异进行加权求和，构建距离函数，以衡量路径之间的相似度。这种方法存在一些局限。如果一条很长的轨迹中只有一小段呈现异常特征，那么这种异常很可能会被整条轨迹的数据平均掉，因而无法识别 [10]。要让异常值显现出来，数据集中大多数轨迹必须具有相似的移动行为。然而，在许多实际场景中，即使 GPS轨迹数据不是异常值，其移动行为也各不相同。此外，该方法会删除整条异常轨迹，造成不必要的数据丢失。

一些研究并不识别和删除异常轨迹，而是关注轨迹中的异常值片段。这些研究采用分割检测模型 [10-15]。分割检测模型首先对轨迹进行分割，然后删除异常子轨迹。研究者以邻近轨迹为参照，从位置和角度等方面检测异常子轨迹。这类方法的一个主要局限是，异常值检测依赖邻近轨迹。如果没有邻近轨迹，独立的子轨迹就会被判定为异常值并因此被删除。这种方法要求邻近轨迹具有较高的空间密度；否则，正常子轨迹就会被误判为异常值，导致数据丢失过多。

此外，还有基于密度的异常值检测方法，最初是通过定义局部异常因子（LOF）[16]提出的。LOF用于衡量一个点与其邻近点的孤立程度。LOF较高的点会被判定为异常值。邻域大小对这种方法至关重要：如果邻域太小，可能无法识别异常值；如果太大，则需要付出过高的计算成本。事实证明，邻域大小对结果非常敏感 [17]。LOF需要对邻近点进行大量计算，因此不适用于大规模（例如全州或全国范围）的数据集。被动采集的GPS数据有时频率较低，因此出现与其他点在地理位置上相距甚远的孤立点是正常的。这些点并非异常值，也可能是行程中的中间点。然而，基于密度的方法可能会错误地将其删除。

与基于密度的方法类似，一些基于聚类的方法[18, 19] 根据空间密度检测异常值。然而，基于聚类的方法最初旨在生成簇，而非识别异常值。无法形成簇或无法归入任何簇的点会被识别为异常值，但在其他研究背景下，这些点实际上可能是正常点。被动采集的GPS数据在不同位置可能呈现不同的空间密度，因此簇大小难以确定。此外，无法形成簇的低密度点不一定是异常值；在基于阈值的行程识别研究等其他情境中，这些点反而可能至关重要。基于分类的方法 [20] 通过利用高质量的训练数据集来检测异常轨迹，但被动采集的GPS数据中并没有这样的数据集。

另一类检测数据振荡的方法是基于模式的方法。该方法最初由 [21] 提出，用于识别蜂窝数据中的乒乓切换（移动设备连接的蜂窝基站会在短时间内发生变化，导致其地理位置快速来回变动）。该方法根据蜂窝基站序列定义了数据振荡的两种移动模式。后续的一些研究对该方法进行了改进，以不同方式识别乒乓现象，例如额外考虑时空特征 [7], 、建筑物聚类 [3], ，并定义不同的时间窗口 [2]。所有这些方法都是专门为蜂窝基站数据或蜂窝数据中的乒乓现象设计的。GPS数据的位置精度和准确度高于蜂窝数据，因此，GPS移动更加细粒度且不确定，导致异常移动模式难以定义。被动采集的GPS数据通常具有不同的采样频率，因此其中的数据振荡呈现出比乒乓振荡更加多样、复杂的模式。尽管基于模式的方法并不充分，也不适合直接应用于被动采集的GPS数据，但作者看到了这类方法的优点。本文提出了若干启发式方法，通过修订和改进一种先进的基于模式的方法，来处理被动采集的GPS数据中的振荡。

2. 数据振荡识别

如前所述，基于模式的方法对于开发从被动采集的GPS数据中识别数据振荡（即异常点）的方法，具有最大的参考价值。尽管这些方法专为蜂窝数据中的乒乓现象设计，但其核心思想颇具启发性：数据振荡通常表现为异常移动模式和不可能出现的高移动速度 [1 - 3, 7, 21]。最先进的基于模式的方法 [1] 设计了若干启发式方法，以从蜂窝数据中识别数据振荡，并在此基础上进行修订和改进，使其适用于被动采集的GPS数据。此外，本文还融入了递归前瞻滤波器 [1 的递归思想，以及如前所述对后续第三个点进行额外检查的做法。

本研究从五个方面进行了修订和改进。

- 首先，通过使用自适应动态参数而非静态参数，构建了灵活的时空公式，其中涵盖了更多数据振荡情形。• 其次，通过使用 Geohash 分区替代原始纬度和经度或建筑物聚类，降低了位置不确定性，节省了大量计算成本，在大规模研究中也更具成本效益。• 第三，受“稳定时期”的启发，设计了“稳定社区”和“稳定区域”的定义，用于标记被认为是正常点的 GPS 点；通过构建局部社区，这种方法更加灵活，同时进一步考虑停留时间和频率后也更加可靠。• 第四，通过进一步考虑连续逐一发生的数据振荡，纳入了更多数据振荡的移动模式。• 第五，参数的选择依据敏感性测试或文献综述中的经验，而不是任意确定。为了表示移动模式，研究通常会将原始纬度和经度

位置转换为簇 [3]。为降低计算成本，采用了另一种方法：将纬度和经度坐标投影到 7 级 geohash 区域（赤道附近为 152.9 米 × 152。4 米）。这大幅降低了 GPS 点的位置不确定性，并为构建移动模式提供了基础。因此，每台设备的原始 GPS移动均表示为一系列 7 级 geohash 区域。提出异常移动模式，并据此移除数据振荡。

2.1. 稳定社区与稳定区域

一个重要的假设是：如果某台设备在某个地点被频繁观测到（有多条 GPS 观测记录），或者停留时间足够长（停留时间长），则认为该地点是一次真实访问。此处所说的地点真实访问并非指一个 7 级 geohash 区域。相反，由于设备并非只静止停留在某处，大多数情况下都在移动途中，因此我们构建一个动态社区来表示某个地点。这里的社区是指某台设备在某个时间段内一组局部的连续观测记录。单台设备可以形成多个不同的社区。每个社区通过不断纳入彼此足够接近的连续观测记录而逐渐扩展，也就是说，两条连续观测记录之间的距离小于 $dist_c$ 英里。由于设备轨迹由一系列 7 级 geohash 区域表示，基于这些轨迹构建的社区也由一组 7 级 geohash 区域组成。有时，一个社区可能只包含一个 geohash 区域，其中有一个或多个 GPS 点。$dist_c$ 根据 GPS 数据的统计分析确定。例如，经过测试的 GPS 数据集显示，每台设备两条连续 GPS 观测记录之间的距离均值为 0.4 英里。因此，可将其四舍五入为 0.5 英里，作为构建社区的阈值，以容纳边缘变化。

请注意，社区可能是一组数据振荡。根据对测试数据的调查，当发生一次数据振荡时，第一次振荡周围可能还会出现额外的振荡。为了区分真实社区和振荡社区，本文提出了稳定社区的定义：如果某设备在某个社区内出现的次数足够多，或停留时间足够长，则认为该社区代表一次真实访问，并将其判定为稳定社区。稳定社区所包含的所有 Geohash 分区均定义为稳定区域，而稳定社区内的所有 GPS 点均判定为真实正常点。每个社区有两个属性：频率和持续时间。频率是 GPS 点的总数，持续时间是第一个点与最后一个点之间的时间间隔。频率阈值可根据对 GPS 点的分析确定。例如，测试数据显示，每个 7 级 geohash 区域中的 GPS 点平均数量为五个。持续时间阈值可设为 300 秒，因为这是根据 GPS 数据确定行程停留点时常用的停留时间值。对测试数据进行的敏感性分析（图 1）还显示，当持续时间阈值设为 300 秒时，识别出的振荡数量达到峰值 1.48。

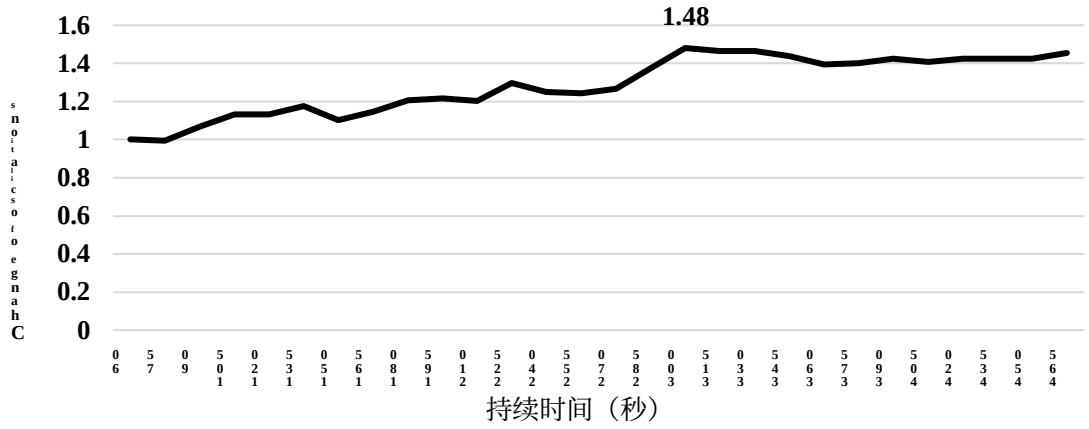


图 1. 持续时间阈值的敏感性分析

图2展示了某设备的三个社区，其中两个是稳定的，一个是不稳定的。蓝色社区由两个 7 级 geohash 区域组成，绿色社区由三个 7 级 geohash 区域组成。它们是稳定社区，即被假定为真实访问，因为数据频率较高（例如，GPS 点超过五个）。红色社区是不稳定的，其中仅包含一个 GPS 点。不稳定并不意味着存在振荡；这只是表示在检测附近振荡时，它们不再具备被视为真实正常访问的资格。

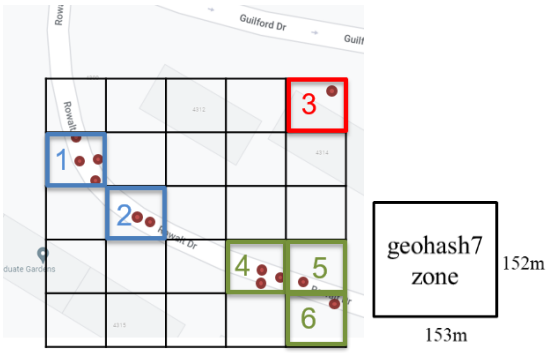


图2. 稳定与不稳定社区示例

2.2. 振荡识别启发式方法

根据稳定社区和稳定区域的定义，我们提出了两种用于 7 级 geohash 区域层级的启发式方法（启发式方法1a、1b）和两种用于社区层级的启发式方法（启发式方法2a、2b），以识别数据振荡。这些方法旨在递归地应用于 GPS数据。详情如下。

- 启发式方法1a：假设设备不可能在极短的时间间隔内完成往返行程，也不可能通过地面交通以异常高的速度移动到遥远的地方。例如，如果一台设备在30秒内从一个处于稳定状态的稳定7级 geohash 区域出发，又返回同一个处于稳定状态的7级 geohash 区域，则判定所有中间观测记录均为数据振荡，并将其移除。时间阈值 t_{min} 按以下公式计算：$t_{min} = 2 * dist_c/v_{max}$。当 $dist_c$ = 0英里且 $v_{max} = 155/1.3 = 120\ mp$小时，$t_{min}$ = 30 秒。使用 $dist_c$ 是因为，当中间的 geohash7 区域是不稳定区域时，它们必须至少相距 $dist_c$ 。否则，它与之前所在的社区属于同一个稳定社区。每小时155英里是允许驾驶速度 [9]的两倍。由于公式中使用的是大地测量距离，而非道路网络距离，因此采用1.3的绕行系数来进行调整

差异。图3展示了启发式方法1a的一个例子：设备从顶部的一个稳定7级 geohash 区域出发，随后跳到右下方的另一个区域，再跳回起始区域。右侧的中间点被判定为数据振荡。

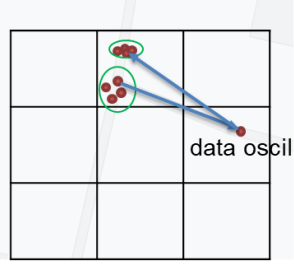


图3. 启发式方法1a示例

- 启发式方法1b：考虑到被动采集的GPS数据频率各异，有时频率可能较低，因此要满足上述检测标准可能要求过高。因此，我们进行了补充检查。我们假设设备无法通过地面交通以异常高的速度移动到遥远的地方。对于任意一对连续的7 级 geohash 区域，如果其中一个区域稳定而另一个区域不稳定，则会检查不稳定区域，查看是否存在振荡。如果这两个7 级 geohash 区域之间的距离超过$dist_g$（例如，5英里），且两区域之间的时间间隔少于t_g（例如，2.5分钟），则判定不稳定区域内的所有观测记录均为数据振荡，并将其删除。我们不仅使用单一速度参数，还会考虑距离和时间间隔。这是因为漂移GPS点很常见，可能显示出非常高的速度，但实际上并非真正的数据振荡。这些漂移点作为数据振荡，不会对推断产生不良影响。此启发式方法以蜂窝数据为例[1]，使用5公里和1分钟。这里通过应用7 级 geohash 区域对启发式方法进行了修订，并采用调整后的距离阈值5英里和时间阈值2.5分钟，对应速度为120英里/小时（考虑绕行系数后的最高允许速度）。阈值可根据具体情况调整。图4展示了启发式方法1b的一个示例。设备形成一个稳定的7 级 geohash 区域（sg2），并检查其前一个区域（g1）和后一个区域（g3）是否可能为异常值。当满足两个阈值（$dist_g$、 t_g）时，即可识别数据振荡，例如g3中包含的点。

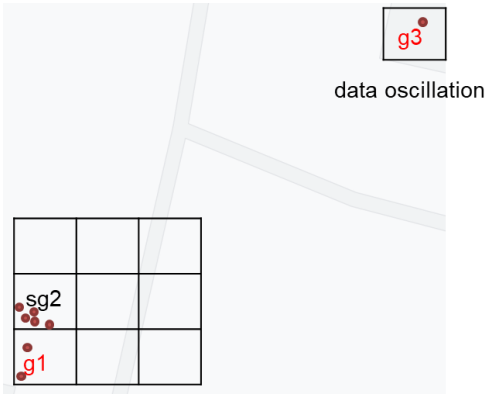


图4. 启发式方法1b示例

- 启发式方法2a：这里考虑的振荡，是从几何角度看显而易见的振荡。设备中任意连续三个社区都可以构成一个三角形。根据以下两个条件，检查社区之间的距离和速度。当两个条件都满足时，中间社区中的所有 GPS 点很可能是振荡，因此会被移除。

$$v_{i-1,i} * v_{i,i+1} > 155 * 155$$
$$dist_{i-1,i+1} < 0.25 * \min(dist_{i-1,i}, dist_{i,i+1})$$

其中 $v_{i-1,\ i}$ 和 $dist_{i-1,\ i}$ 分别表示从第一个社区到第二个社区的速度和距离；同样，$v_{i,\ i+1}$ 和 $dist_{i,\ i+1}$ 分别表示从第二个社区到第三个社区的速度和距离；$dist_{i-1,\ i+1}$ 表示第一个社区与第三个社区之间的距离。如图 5 所示，如果两个条件都满足，则确定中间社区（C2）中包含的点属于数据振荡。

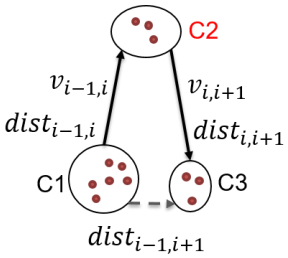


图 5. 启发式方法 2a 示例

- 启发式方法 2b：有时，启发式方法 2a 所述的情况可能连续出现，即一组社区中任意连续的三个社区都满足启发式方法 2a 的两个条件。在这种情况下，需要结合其他因素来判断是否存在真实振荡。从第一个社区开始，按顺序为所有社区标注位置编号（即 1、2、3.....）。因此，社区分为两组：位置为奇数的社区和位置为偶数的社区。假设平均停留时间较长的社区代表真实正常访问，并予以保留。如图 6 所示，连续观察到一组三个社区，例如 (C1, C2, C3)、(C2, C3, C4)、.....、(C6, C7, C8)，均满足启发式方法 2a 的两个条件。然后，根据各组社区的平均停留时间，保留位置为偶数或奇数的那一组。

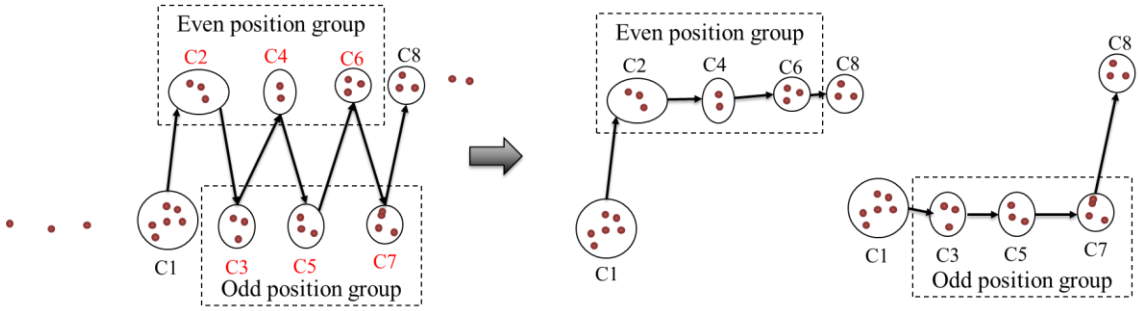


图 6。启发式方法 2b 示例

3. 讨论

GPS数据广泛应用于各种交通研究中，但数据预处理，尤其是移除振荡点，却常被忽视，这正是本文的研究缘起。作者全面回顾了以往关于位置数据（即 GPS 点和轨迹数据、蜂窝数据及蜂窝基站数据）中数据振荡的研究，并探讨了被动采集的GPS数据的优势、局限性和适用性。考虑到被动采集的GPS数据中振荡的特点（即不确定性更高、更多样且更复杂），作者借鉴了以往研究中的一些优点，并对其作出了若干修订和改进。据作者所知，利用 Geohash分区降低位置不确定性并定义移动模式的方法尚属首次探讨。与使用原始经纬度或簇的其他方法相比，该方法不仅计算成本更低，而且由于地理哈希分区的级别可以

调整，因此对各种案例研究也表现出很高的灵活性和适应性。对于高频 GPS数据，可以采用级别更高的Geohash分区系统（例如，赤道附近尺寸为 38.2 米 x 19 米的 8 级 Geohash分区）。此外，本文基于稳定社区和稳定 Geohash分区这一创新理念，通过两级启发式方法纳入了被动采集的GPS数据中各种振荡模式。与许多现有研究仅依据速度移除GPS数据中的振荡位置点不同，本文首先对正常点作出假设，然后依据其他点与正常点之间的时空关系，判断这些点是否属于振荡。考虑到振荡在被动采集的GPS数据中可能呈现不同尺度，本文采用了动态参数。尽管数据振荡识别方法已相当成熟，但这些方法是为蜂窝数据设计的，无法处理被动采集的GPS数据中的振荡。其他与GPS相关的振荡识别方法也各有局限：基于速度的方法可能会将正常点误判为异常值；基于密度和基于聚类的方法可能会将孤立点误判为异常值，而孤立点在被动采集的GPS数据中可能普遍存在；基于距离的方法和分区检测法以空间密度高的数据集为前提，应用于低密度数据集时可能导致大量数据丢失。本文提出的启发式方法旨在谨慎识别振荡，适用于不同地点数据密度各异的被动采集GPS数据，具有实际意义。

致谢 我们谨向.....致谢并

感谢并致意本项目的合作伙伴和数据来源：(1) 美国交通部联邦公路管理局提供的部分资金支持；(2) 亚马逊云科技及其高级解决方案架构师徐建军和 Greg Grieff 提供的云计算和技术支持。

作者贡献 作者确认其贡献

对论文的贡献如下：研究构思与设计：Q. Sun、Y. Pan、A. Darzi；数据收集与准备：Q. Sun；结果分析与阐释：Q. Sun、Y. Pan、A. Darzi；论文初稿撰写：Q. Sun、Y. Pan、A. Darzi。所有作者均审阅了研究结果，并批准了论文最终版本。

参考文献 1. Wu, W., Wan

g, Y., Gomes, J. B., Anh, D. T., Antonatos, S., Xue, M., Yang, P., Yap, G. E., Li, X., Krishnaswamy, S., Decraene, J., & Shi-Nash, A. (2014年). 用于移动性建模的手机蜂窝塔数据振荡消解. 2014年IEEE第15届移动数据管理国际会议. https://doi.org/10.1109/mdm.2014.46 2. Qi, L., Qiao, Y., Abdesslem, F. B., Ma, Z., & Yang, J. (2016年). 海量手机流量数据振荡消解. 首届移动数据研讨会论文集——MobiData' 16. https://doi.org/10.1145/2935755.2935759 3. Wang, F., & Chen, C. (2018年). 从被动生成的手机数据中提取移动模式所需的数据处理. 《交通研究C辑：新兴技术》, 87, 58–74. https://doi.org/10.1016/j.trc.2017.12.003 4. Hawkins, D. (1980年). 离群值识别. Chapman & Hall, 伦敦. 5. Barnett, V., Lewis, T. (1994年). 统计数据中的异常值. John Wiley & Sons, 奇切斯特 . 6. Knorr, E. M., Ng, R. T., & Tucakov, V. (2000 年). 基于距离的离群值：算法与应用. 《VLDB期刊：超大型数据库国际期刊》, 8(3–4), 237–253. https://doi.org/10.1007/s007780050006 7. Iovan, C., Olteanu-Raimond, A. M., Couronné, T., & Smoreda, Z. (2013 年). 移动与通话：人类移动性研究中的手机数据质量测量与时空不确定性 . 《欧洲核心的地理信息科学》, 247–265. https://doi.org/10.1007/978-3-319-00615-4_14 8. Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S., & Eriksson, J. (2009年). VTrack. 第7届ACM嵌入式网络传感器系统会议论文集——SenSys '09 1644048 9 10 1145 1644038 2014年 . https://doi.org/ . / . . Horn, C., Klampfl, S., Cik, M., & Reiter, T. ( ). 检测手机数据中的离群值. 《交通研究记录：交通研究委员会期刊》, 2405(1), 49–56. https://doi.org/10.3141/2405-07 10. J. Lee, J. Han and X. Li. (2008年). 轨迹离群检测：分区检测框架. 2008年IEEE第24届数据工程国际会议, 140-149. doi: 10.1109/ICDE.2008.4497422. 11. Yang, X., & Tang, L. (2016年). 众包大规模轨迹数据过滤：分区过滤模型. ISPRS - 国际摄影测量、遥感与空间信息科学档案, XLI-B2, 257–262. https://doi.org/10.5194/isprsarchives-xli-b2-257-2016 12. Zhang, L., & Wang, Z. (2011年). 基于时间参考和速度的轨迹划分方法. 《融合信息技术期刊》, 6(8), 134–142. https://doi.org/10.4156/jcit.vol6.issue8.16 13. Krishnan, S., Garg, A., Patil, S., Lea, C., Hager, G., Abbeel, P., & Goldberg, K. (2017年). 过渡状态聚类：用于机器人学习的无监督手术轨迹分割. 《国际机器人研究期刊》, 36(13–14), 1595–1618. https://doi.org/10.1177/0278364917743319 14. Yu, Y., Cao, L., Rundensteiner, E. A., & Wang, Q. (2014年). 在大规模轨迹流中检测移动对象离群值. 第20届ACM SIGKDD知识发现与数据挖掘国际会议论文集. https://doi.org/10.1145/2623330.2623735 15. Gupta, M., Gao, J., Aggarwal, C., & Han, J. (2014年). 时序数据离群检测. 《数据挖掘与知识发现综合讲座》, 5(1), 1–129. https://doi.org/10.2200/s00573ed1v01y201403dmk008 16. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000年). LOF. 《ACM SIGMOD记录》, 29(2), 93–104. https://doi.org/10.1145/335191.335388 17. Papadimitriou, S., Kitagawa, H., Gibbons, P., & Faloutsos, C. (2003年). LOCI：使用局部相关积分快速检测离群值. 第19届数据工程国际会议论文集 (Cat. No.03CH37405). https://doi.org/10.1109/icde.2003.1260802

18. M. Ester, H.-P. Kriegel, J. Sander, X. Xu.（1996）。一种用于在含噪声大型空间数据库中发现簇的基于密度的算法。KDD会议论文集。19. S. Guha, R. Rastogi, K. Shim.（1998）。CURE：一种适用于大型数据库的高效聚类算法。ACM SIGMOD会议论文集。20. Li, X., Han, J., Kim, S., & Gonzalez, H.（2007）。ROAM：海量移动对象数据集中的基于规则和模式的异常检测。2007年SIAM国际数据挖掘会议论文集。 https://doi.org/10.1137/1.9781611972771.2521. Lee, J. K., & Hou, J. C.（2006）。用户移动性的稳态行为和瞬态行为建模：第七届ACM移动自组织网络与计算国际研讨会论文集 - *MobiHoc '06*。 https://doi.org/10.1145/1132905.1132915