第3章 移动群智感知网络感知质量

移动群智感知网络的感知质量包含时空覆盖质量和数据质量两个层面，前者关注是否能采集到足够多的数据，而后者关注数据是否足够准确和可信。由于移动群智感知网络是利用参与用户的移动性来扩展感知覆盖范围的，因此时空覆盖质量与人类的移动模式密切相关，而人类的移动模式是非常复杂的，既有一定的随机性，又存在一些有趣的规律，这种复杂的移动模式使得我们很难对时空覆盖质量进行度量，甚至缺少统一的度量指标。同时，由于参与用户数量和人的移动范围所限，总是存在一些区域在某些时间没有任何用户采集感知数据，即所谓的“感知盲区”问题。另一方面，感知数据的质量也受很多因素的影响，主要因素包括：①用户所使用的感知设备类型，例如价格高昂的高端手机的传感器一般比价格低廉的低端手机的传感器精度高；②用户的数据采集方式，如把手机拿在手里采集环境噪声比把手机放在衣服口袋或手提包里采集环境噪声的数据质量高；③用户的主观认知能力，例如基于移动群智感知的图像搜索应用[1]依赖用户对图像的识别能力，而不同用户对同一图像的认知可能是不一样的；④用户的参与态度，例如有些用户会严格按照要求来采集数据，而有些用户会比较随意，甚至有些恶意用户会为了获得感知平台提供的任务报酬或其他原因而上传虚假、伪造的数据。以上因素都会造成感知数据质量参差不齐，使得感

知数据质量很难保障。

针对上述问题，本章将分别从时空覆盖质量和数据质量两个层面探讨感知质量度量与保障问题以及对应的解决方法，从而指导移动群智感知网络的部署和应用。首先，将在 3.1、3.2、3.3 三节依次介绍三种时空覆盖质量的度量模型与方法。其中，前两节主要面向需要对整个城市的每个区域进行连续监测的大规模城市感知应用，分别从时间维度和空间维度引入“机会覆盖率”和“城市分辨率”两种度量指标；3.3 节则面向“以地点为中心”的移动群智感知应用，介绍“地点覆盖率”度量指标。其次，将在 3.4 和 3.5 两节介绍两种感知质量增强方法来解决“感知盲区”问题。其中，3.4 节介绍基于压缩感知质量增强方法，利用感知数据内在的“可被稀疏表达”特性来恢复数据；而 3.5 节介绍基于多源数据相关性质量增强方法，利用不同来源的感知数据间的相关性来恢复数据。最后，将在 3.6 和 3.7 两节探讨数据质量问题。其中，3.6 节针对数据准确性问题，介绍典型的数据质量度量和保障方法，其基本思路是发挥集体的智慧来抵御个人数据不准确的影响，从而提高整体数据的准确性；而 3.7 节针对数据质量可信性问题，介绍基于信誉系统的数据质量度量和保障方法，其基本思路是评估和记录用户的历史感知数据的可信性，并将其用在未来的系统交互过程中，对于信誉度低的用户感知数据采用的可能性也比较低，同时也会采用相应的激励或惩罚措施。

3.1　机会覆盖率

3.1.1　机会覆盖模型

覆盖是衡量网络服务质量的一个重要性能指标，与参与感知的节点个数密切相关。在实际部署一个感知系统之前，首先需要了解多少个节点能达到怎样的覆盖质量，才能合理规划网络规模。在传统的传感器网络中，覆盖问题一直备受关注。例如，在传统的固定部署传感器网络中，通常需要监测区域内的每个点总是被至少一个传感器节点覆盖，并且覆盖质量不会随着时间而改变[2-4]；在传统的移动传感器网络中，则通常需要监测区域内的每个点在一定时间段内被覆盖，而不是一直被覆盖，即覆盖质量是随着时间而变化的[5-7]。类似地，在利用移动群智感知网络进行城市感知时，达到特定的覆盖质量需求也是必需的。然而，移动群智感知网络中的覆盖不同于传统的传感器网络中的覆盖，它与人移动的机会性密切相关，我们将其称为“机会覆盖”。以城市空气质量监测为例，假定我们计划使用大量的出租车携带空气质量传感器，对北京市的五环内区域进行监测，构建每天 6 点至 24 点时间段的空气质量感知图谱。事实上，这里有两个基本问题有待回答：①怎样度量这些出租车提供的感知机会，以及它们能达到的感知质量？②需要部署多少

辆出租车才能达到所需的感知质量？

针对第一个问题，考虑到感知覆盖的时空变化性，我们提出“覆盖间隔时间”作为度量指标[8]。具体来说，在时间域上将所关注的时间段 T 划分为多个同等大小的采样周期 T_S，如图 3-1(a)所示；同时，在空间域上将整个城市感知区域划分为多个同等大小的网格单元，如图 3-1(b)所示，其中，每个网格单元的大小代表由应用需求决定的空间感知粒度。当一个新的采样周期到来，并且节点的位置恰好在某个网格单元内时，称该网格单元被覆盖一次。“覆盖间隔时间”则定义为每个网格单元被连续覆盖两次的间隔时间，可用来描述每个网格单元被覆盖的机会。如图 3-2 所示，一个网格单元被节点 v_1、v_2 和 v_3 覆盖了三次，其覆盖持续时间分别为 $3T_S$，$2T_S$ 和 T_S。同时，可以从中提取两个覆盖间隔时间(在图中用 T_I 标示)。明显地，覆盖间隔时间直接反映了覆盖质量，即覆盖间隔时间越短，则网格单元的覆盖质量越好。直觉上来说，覆盖间隔时间的分布与网格单元的大小和节点的个数两个因素密切相关。也就是说，网格单元越大，节点个数越多，则覆盖间隔时间越小。为了描述这些因素，将网格单元 g_i 在 n 个节点的条件下的覆盖间隔时间分布表示为式(3-1)：

$$F_i(\tau;n) = P\{T_I \leqslant \tau \mid (g = g_i, N = n)\} \tag{3-1}$$

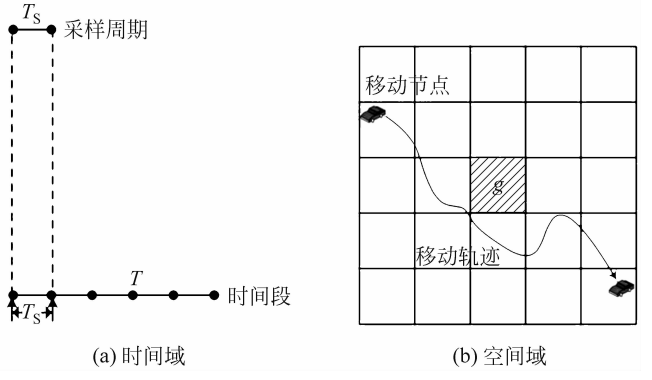


图 3-1 离散化的时空感知域示意图

针对第二个问题，则需要描述整个城市感知区域的覆盖质量与节点个数之间的关系，为此，我们定义了“机会覆盖率”，即在时间间隔 τ 内能被机会覆盖的网格单元占所有网格单元的比例的期望值，可表达为式(3-2)：

$$f_I(\tau) = \frac{\sum_{i=1}^{m} F_i(\tau;n)}{m} \tag{3-2}$$

从上述表达式可以看出，机会覆盖率与节点个数和时间间隔呈单调递增关系。

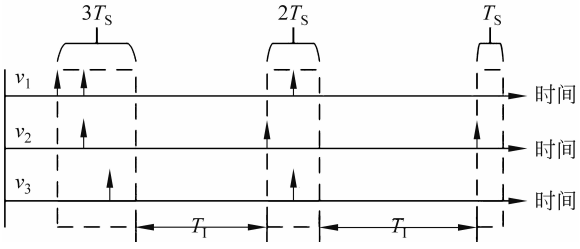


图 3-2　覆盖间隔时间示意图

3.1.2　覆盖间隔时间分布规律

覆盖间隔时间与机会网络中影响数据包投递性能的一个重要因素“接触间隔时间”(inter-contact time)十分相似。最近,一些基于人或车辆的移动轨迹数据的实证研究表明,节点的接触间隔时间通常服从截断的幂律分布或指数分布[9-15]。事实上,许多研究已经表明人类出行的时空特征(如出行间隔时间、行走距离等)、通信模式(如发邮件和打电话的间隔时间)、工作和行为模式(如网页访问间隔时间、视频点播间隔时间)等均服从某种特定的规律[16,17],一般表现为指数分布、幂律分布、截断的幂律分布三种形式,如表 3-1 所示。其中,截断的帕累托分布在头部呈现幂律分布趋势,在尾部则呈现指数衰退趋势。因此,我们猜想覆盖间隔时间也应该服从某种特定规律。

表 3-1　三种统计模型的描述

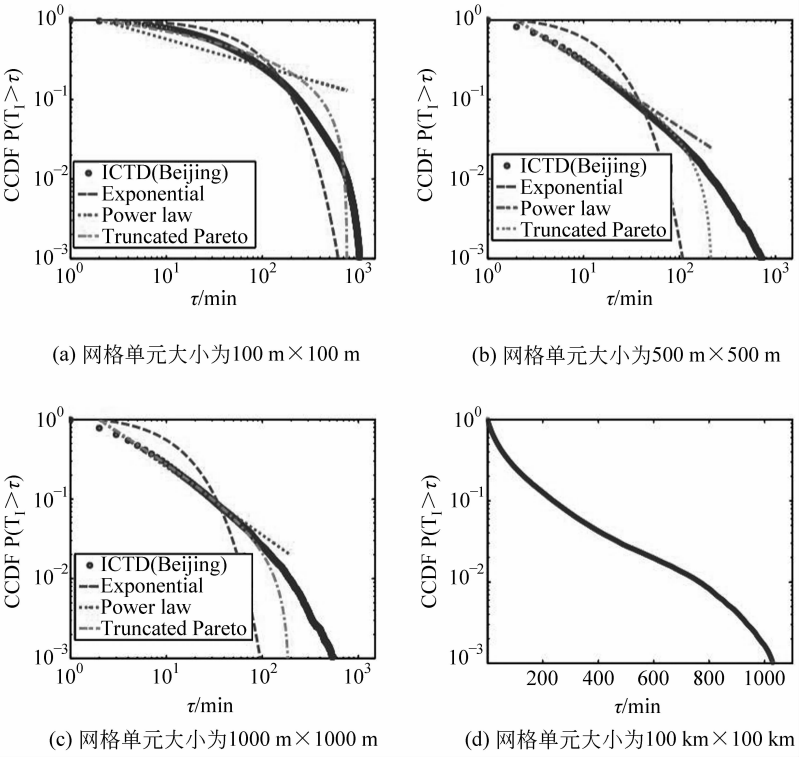
分布类型(数据范围)	概率密度函数(PDF)$f(x)$	累积分布函数(CDF)$F(x)$
指数分布($x\geqslant a$)	$\lambda e^{-\lambda(x-a)}$	$1-e^{-\lambda(x-a)}$
幂律分布($x\geqslant a$)	$(\lambda-1)a^{\lambda-1}x^{-\lambda}$	$1-a^{\lambda-1}x^{1-\lambda}$
截断的帕累托分布($a\leqslant x\leqslant b$)	$\frac{\lambda a^{\lambda}x^{-\lambda-1}}{1-(a/b)^{\lambda}}$	$1-\frac{a^{\lambda}(x^{-\lambda}-b^{-\lambda})}{1-(a/b)^{\lambda}}$

为了分析覆盖间隔时间分布的规律,我们使用两个出租车移动轨迹数据集。第一个数据集包含北京市从 2008 年 2 月 2 日至 2 月 8 日期间的 10 357 辆出租车的 GPS 移动轨迹数据,其平均采样间隔为约 177s[18]。第二个数据集包含上海市在 2007 年 2 月 20 日一天内 4316 辆出租车的 GPS 移动轨迹数据,其平均采样间隔在有乘客时为约 60s,在没有乘客时为约 20s[15]。经过预处理获得北京市 4067 辆出租车(2008 年 2 月 3 日)和上海市 2079 辆出租车从 6:00 至 24:00 期间每 60s 的 GPS 位置数据两个新的移动轨迹数据集,如图 3-3 所示。我们选择北京市五环(面积约 900km^2)内和上海市相同面积区域内的移动轨迹数据进行分析,时间段 T 为从 6:00 至 24:00 的 18h,采样周期 T_S 为 60s。分别考虑不同的网格单元大小和节点个数,利用 Akaike 测试[19,20]对这两个数据集进行实证分析,得到相似的结

果：不论网格单元是多大，也不论节点个数是多少，覆盖间隔时间分布均最符合截断的帕累托分布。图 3-4 和图 3-5 分别显示不同网格单元大小情况下（出租车个数取 1000）和不同节点个数情况下（网格单元取 100m×100m）的北京数据集的覆盖间隔时间分布。

(a) 北京市面积约900 km²的五环区域　　(b) 上海市面积约900 km²的区域

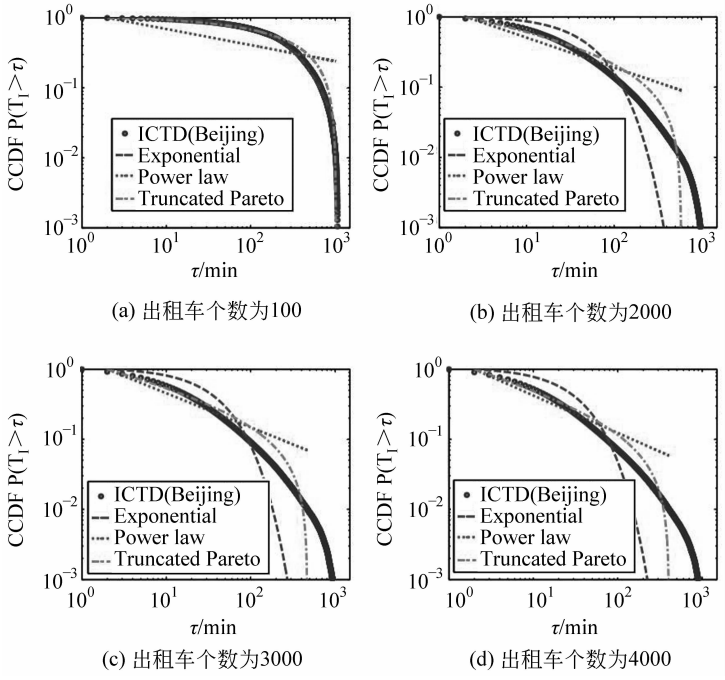
图 3-3　出租车移动轨迹分布



(a) 网格单元大小为100 m×100 m　　(b) 网格单元大小为500 m×500 m

(c) 网格单元大小为1000 m×1000 m　　(d) 网格单元大小为100 km×100 km

图 3-4　不同网格单元大小情况下北京数据集的覆盖间隔时间分布



(a) 出租车个数为100　(b) 出租车个数为2000

(c) 出租车个数为3000　(d) 出租车个数为4000

图 3-5　不同出租车个数情况下北京数据集的覆盖间隔时间分布

3.1.3　城市机会覆盖率分析

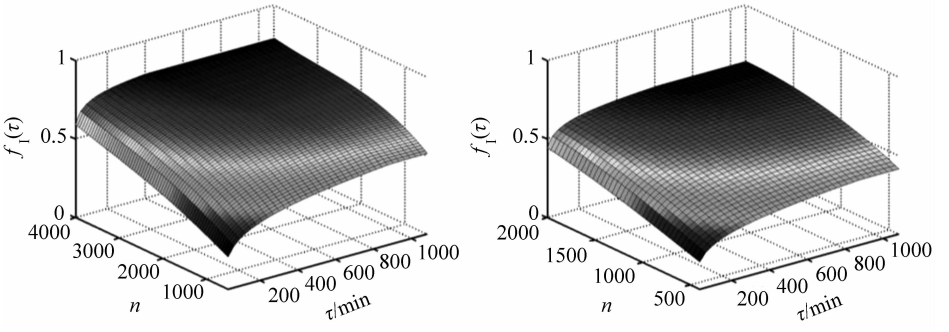
由 3.1.1 节可知，必须首先获取每个网格单元的覆盖间隔时间分布 $F_i(\tau;n)$，然后推导出整个城市机会覆盖率 $f_1(\tau)$。然而，以下两个原因导致难以直接计算：①存在一些很少有节点访问的网格单元，其覆盖间隔时间数据较少，因而难以准确估计其分布；②网格单元个数太多，例如，当网格单元大小为 100m×100m 时，北京五环内一共有 9 万个网格单元，因而对所有网格单元的覆盖间隔时间进行拟合的计算代价较高。因此，我们基于所有网格单元的覆盖间隔时间分布，以简单有效的方式对整个城市的机会覆盖率进行分析。将整个时间段 T 内能至少被机会覆盖一次的网格单元的比例表示为 $p(n)$，与 n 呈单调递增关系。那么，整个城市的机会覆盖率可表达为如下的式(3-3)：

$$f_1(\tau) = \frac{F_a(\tau;n) \times m \times p(n)}{m} = F_a(\tau;n) \times p(n) \tag{3-3}$$

其中，$F_a(\tau;n)$是所有网格单元的聚集的覆盖间隔时间累积分布函数，其服从如式(3-4)所示截断的帕累托分布：

$$F_a(\tau;n) = 1 - \frac{a^{\lambda}(\tau^{-\lambda} - b^{-\lambda})}{1 - (a/b)^{\lambda}} \tag{3-4}$$

接下来，分别提取不同个数的出租车的移动轨迹，分析覆盖间隔时间分布与节点个数之间的关系，除了覆盖间隔时间都服从截断的帕累托分布之外，我们还使用最小二乘法对截断的帕累托分布的指数与出租车个数之间的关系进行线性拟合，结果得到很好的拟合结果，所有拟合结果的确定系数都大于 0.98。因此，可以使用线性函数表达 λ 与 n 之间的关系。类似地，我们使用线性函数对 $p(n)$ 也得到很好的拟合结果。通过联合这些线性表达式和公式(3-3)，则可以容易得到 $f_1(\tau)$ 的通用表达式。基于北京和上海两个数据集的最终数值结果如图 3-6 所示。可以明显看到，$f_1(\tau)$ 与 n 和 τ 呈单调递增关系。因此，可以容易估计出达到指定的覆盖质量所需要的节点个数。例如，我们需要分别在北京和上海至少部署 1700 辆和 1900 辆出租车，才能保证其在一个小时的时间间隔内机会覆盖率不小于 50%。尽管不同城市可能需要不同的节点个数以满足指定的机会覆盖率，但是我们提出的模型和方法可以对网络规划问题提供一般性的指导。



(a) 基于北京数据集的结果　　(b) 基于上海数据集的结果

图 3-6　机会覆盖率 $f_1(\tau)$ 与 n 和 τ 的关系

从直觉上来说，人们在不同日期(如工作日、休息日、节日等)总是有不同的移动模式。因此，调查整个城市区域的机会覆盖率的稳定性是很重要的。接下来，我们分析北京市在一周内不同日期(2008 年 2 月 4 日至 2 月 7 日)的机会覆盖率的变化。首先，通过对这 4 天的数据分别进行预处理，获得不同数目出租车从 6:00 至 24:00 期间每 60s 的 GPS 位置数据。不同日期的移动轨迹数据中包含的出租车个数如表 3-2 所示。我们从每一天的数据中随机选择三组出租车(每组有 1700 辆出租车)，并且对这些出租车所能达到的机会覆盖率进行统计分析，结果如表 3-2 所示。可以看出，2 月 4 日和 2 月 5 日的平均机会覆盖率分别达到 50.63%和 48.26%，都非常接近 50%，而 2 月 6 日和 2 月 7 日的机会覆盖率则小于 50%，甚至 2 月 7 日的机会覆盖率仅有 30.84%。但需要注意的是，2 月 7 日恰好是 2008

年中国的春节。众所周知，在中国的春节和除夕两天，大部分中国人选择待在家中团聚而较少外出，因而相比往常出租车顾客更少且在更多时间内保持静止。因此，这可能是这两天机会覆盖率减少的原因。同时，这种现象也暗示我们，在对城市机会感知应用进行网络规划时，应该考虑人的移动模式在不同日期的稳定性。

表 3-2　不同日期的机会覆盖率

日　期	出租车总个数	机会覆盖率			
		第一组	第二组	第三组	平均值
2008 年 2 月 4 日	3982	50.00%	51.34%	50.56%	50.63%
2008 年 2 月 5 日	3727	47.88%	49.13%	47.76%	48.26%
2008 年 2 月 6 日	3549	35.66%	35.95%	35.38%	35.66%
2008 年 2 月 7 日	3600	31.17%	31.24%	30.10%	30.84%

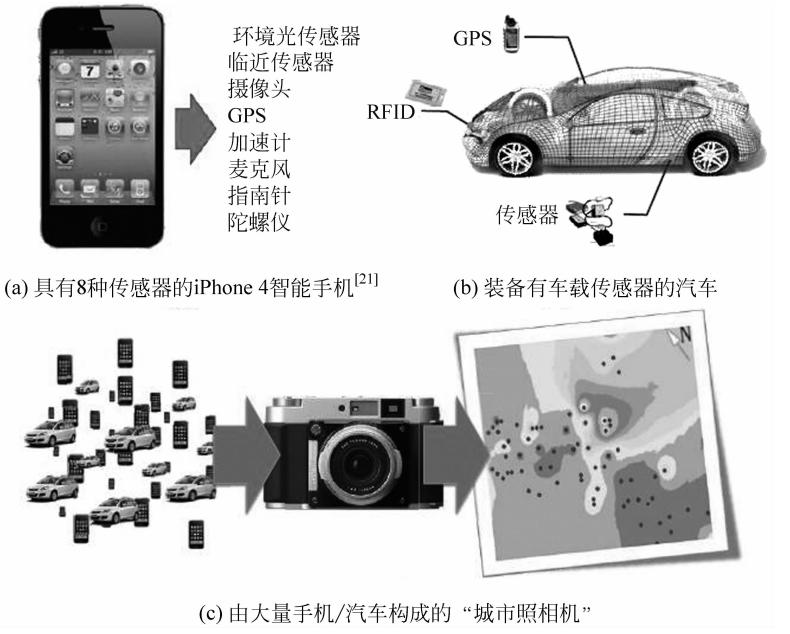
另一方面，从表 3-2 还可以看到，在同一日期具有相同个数出租车的不同组可以达到几乎相同的机会覆盖率（最多有 1.34%的差别）。该现象表明机会覆盖率与节点个数之间的关系是稳定的，而且我们所提出的覆盖度量模型和方法可以准确地估计它们之间的关系。

3.2　城市分辨率

3.2.1　城市分辨率的概念

3.1 节主要是从时间维度考虑感知质量，接下来将重点从空间维度考虑感知质量。一个监测区域的某种环境质量（如 PM2.5 浓度、二氧化碳浓度、噪音级别等）分布情况可以抽象地表示为一个二维信号，类似于一幅图像。在基于移动群智感知的环境监测应用中，移动感知节点将采集到的感知数据发往数据中心，数据中心汇集现有数据，利用感知数据的空间相关性，通过空间插值技术估算感知节点周边的未知数据，从而得到一副完整的感知图谱，相当于生成一副感知图像。如图 3-7 所示，移动群智感知网络就像一台“城市照相机”，而每个移动感知节点就相当于这台照相机的每个“像素”。在传统的图像系统中，分辨率是度量图像质量的一个重要指标。从这个概念得到启发，我们是不是也可以用分辨率来度量感知图像的质量呢？分辨率越高，则代表所部署的移动群智感知网络越能准确地捕获到监测区域内环境现象的变化，这也是对监测区域空间覆盖质量的一种间接度量。然而，与传统数字图像系统中的分辨率定义不同，我们不能简单地将像素数（即移动感知节点个数）看作移动群智感知网络的分辨率。这是因为，我们通常所说的数

字照相机的像素会形成一个精细的网格，而这里所讲的“城市照相机”的像素在城市中则呈现分散的动态化分布。



(a) 具有8种传感器的iPhone 4智能手机[21]　(b) 装备有车载传感器的汽车

(c) 由大量手机/汽车构成的“城市照相机”

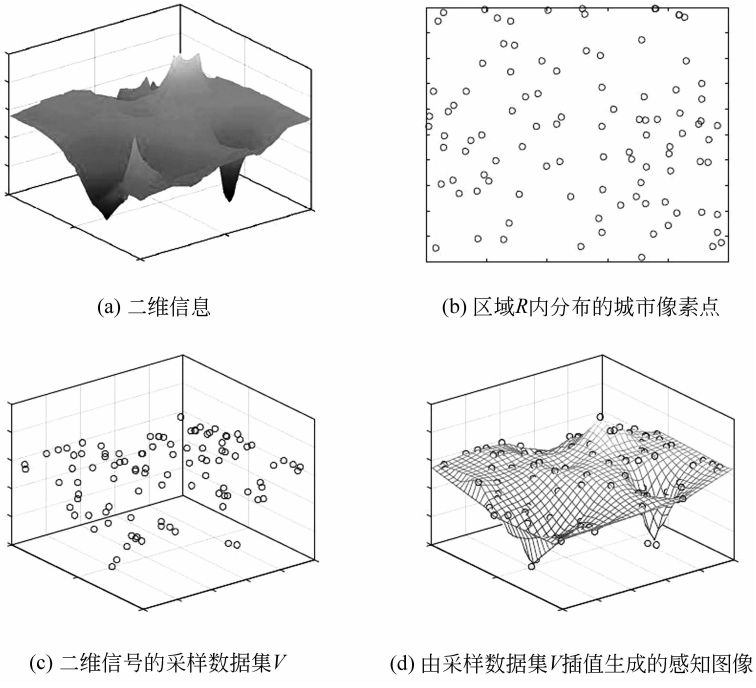
图 3-7　基于移动感知设备的城市感知示意图

针对上述问题，我们提出了一个新的度量指标，称之为“城市分辨率”(urban resolution)，用来描述感知图像的质量(即空间覆盖质量)[22]。下面首先给出相关基本概念的形式化描述，并且介绍感知图像的生成过程，进而阐述城市分辨率的定义及其对感知图像的度量方法。

定义 3-1(城市像素)　具有感知和通信能力的手机/车辆参与城市感知，并且将感知数据上传，我们将这类节点称为城市像素，表示为 u_i。u_i 上传的数据表示为 $v_i=(x_i,y_i,z_i)$，其中，(x_i,y_i)和 z_i 分别表示感知位置和感知数值。

如图 3-8 所示，城市中的环境质量可以被表述成一个二维的连续空间信号。从数学角度讲，这个信号是一个二元函数，函数自变量表示空间位置(x,y)，函数值 $z=I(x,y)$则为环境质量的真实值。假定 s 个城市像素动态地分布在区域 R 中。用 $U=\{u_i,0<i\leqslant s\}$ 表示 t 时刻区域 R 中城市像素的集合，$V=\{v_i,0<i\leqslant s\}$表示该时刻城市像素上传的感知数据集。换句话说，集合 V 就是对函数 $I(x,y)$的一组采样。而生成 t 时刻感知图像的过程，则是通过集合 V 估算区域 R 内所有位置的环境数据。实际上，这是一个二维数据插值问题，也称为空间插值，这种插值方法利用了空间位置上相邻数据的相关性。为了实现空间插值，二

维连续空间信号需要转化成二维离散空间信号，即将区域 R 均匀分割成 $m \times m$ 的网格。至此，下面给出离散化后感知图像的定义。



(a) 二维信息　(b) 区域R内分布的城市像素点

(c) 二维信号的采样数据集V　(d) 由采样数据集V插值生成的感知图像

图 3-8　城市中的环境质量感知图像示意图

定义 3-2(感知图像)　由城市像素集 U 生成的 t 时刻感知图像 Z'，被定义为如下二维矩阵：

$$Z' = \begin{bmatrix} z'_{0,0} & z'_{0,1} & \cdots & z'_{0,m-1} \\ z'_{1,0} & z'_{1,1} & \cdots & z'_{1,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ z'_{m-1,0} & z'_{m-1,1} & \cdots & z'_{m-1,m-1} \end{bmatrix}_{m \times m} \tag{3-5}$$

这里 Z'矩阵中的数值，是根据 U 所对应的感知数据集 V，利用空间插值法(如 Delaunay 三角剖分法[23])计算得出的。

如上所述，二维离散空间环境质量也可以表示成 $m \times m$ 的矩阵形式，这里环境信号矩阵用 Z 表示。因此，通过计算 Z 与 Z'的相关系数，可以得出 Z 与 Z'的相似程度，即感知图像的质量。

定义 3-3(感知图像的质量(QoI))　感知图像的质量 $Q(Z')$，是感知图像 Z'与二维离散空间目标信号 Z 的相似度。这里用相关系数描述其相似度，如式(3-6)所示：

$$Q(Z')\triangleq C(Z',Z)=\frac{\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}(z_{j,k}-\bar{z}_{j,k})(z'_{j,k}-\bar{z}'_{j,k})}{\sqrt{\left(\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}(z_{j,k}-\bar{z}_{j,k})^2\right)\left(\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}(z'_{j,k}-\bar{z}'_{j,k})^2\right)}} \tag{3-6}$$

显然较大的相关系数值，意味着 Z 与 Z' 有较高的相似度，即感知图像 Z' 的质量高。然而，现实场景中，一般无法事先获得原始信号的真实情况 Z，因此，我们无法直接利用上述公式计算 $Q(Z')$。

从另一个角度考虑，如果可以用网格分布形式的感知节点生成一副图像 Z''，使其 QoI 与感知图像 Z' 相近，则网格分布的节点数 $n\times n$ 可以作为感知图像质量的一种间接度量。至此，下面给出城市分辨率的定义。

定义 3-4（城市分辨率） 具有 s 个城市像素的移动群智感知网络的城市分辨率 r，可以表示为 $n_l\times n_l$，这里

$$n_l=\arg\max_{n} C(Z'',Z') \tag{3-7}$$

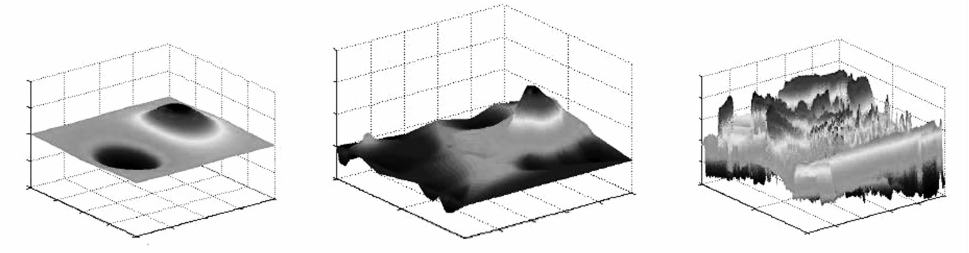
式(3-7)中，Z' 表示感知图像，Z'' 表示由 $n\times n$ 个网格分布的节点生成的感知图像。显然，当 n_l 足够大的时候，Z'' 可以非常接近原始图像 Z，即，$C(Z'',Z)\approx 1$。这意味着感知图像 Z' 的质量很高，移动群智感知网络对监测区域的空间覆盖质量也很高。

3.2.2 城市分辨率变化规律

直觉上说，城市分辨率的变化会受到三方面因素的影响：①原始信号的波动情况；②移动群智感知网络的规模（即城市像素的数量）；③城市像素的分布情况。为了简化问题，本节首先考虑前两种因素对城市分辨率的影响，再结合城市像素的分布特性，最后得出城市分辨率的变化规律。

如图 3-9 所示，我们分别使用三个信号作为原始信号，其中，图 3-9(a)的信号较为平滑，图 3-9(c)中的信号波动最为剧烈。现实中环境信号的波动情况则介于(a)和(c)之间，如图 3-9(b)所示，该图像是由真实的二氧化碳监测数据[24]通过空间插值法得到的感知图像。

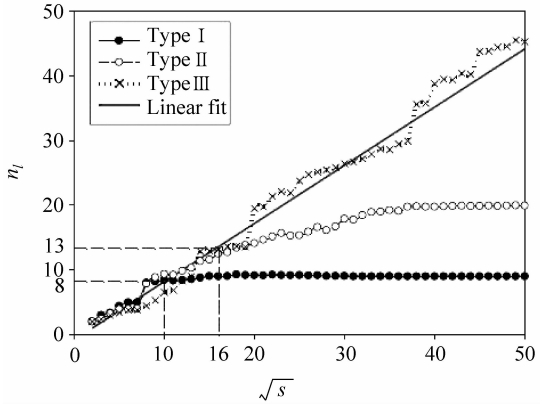
由城市分辨率的定义得知，分辨率 r 的计算并非由解析形式得出，因此我们使用蒙特卡洛随机模拟的方法，统计分析分辨率 r 与城市像素数 s 之间的关系。具体而言，我们在区域 R 内随机生成 s 个城市像素分别对图 3-9(a)(b)(c)中的信号采样。由于城市分辨率 r 被表示成了 n_l 平方的形式。为方便起见，这里考虑了 n_l 与 $\sqrt{s}$ 之间的变化规律。令 $\sqrt{s}$ 的值从 2 变化到 50，间隔为 1。对于每个 $\sqrt{s}$ 的值，随机



(a) 平滑信号 $z=x^{-(\frac{x}{2})^2-(\frac{y}{2})^2-0.5}$　(b) CO_2浓度感知信号[24]　(c) 剧烈波动信号（此信号为Lena图像的灰度信息）

图 3-9　三个不同的原始信号的波动情况

生成 100 副城市像素的位置拓扑，并采样信号得出 n_l 值，之后计算 100 个 n_l 的平均值，也就是该$\sqrt{s}$值所对应的 n_l。计算结果如图 3-10 所示，图 3-9(a)(b)(c)中的信号分别用 Type Ⅰ、Type Ⅱ、Type Ⅲ表示。



注：x 轴为$\sqrt{s}$，y 轴为 n_l，即$\sqrt{r}$

图 3-10　城市分辨率 r 与城市像素数 s 之间的关系

分析图 3-10 中的结果可以看到，n_l 值随着$\sqrt{s}$的增加而增加，这意味着对于三种类型的信号而言，网络规模的扩大都会带来感知图像分辨率的提升，增加了网络对监测区域的空间覆盖程度。与此同时，我们还发现，三类信号 n_l 随$\sqrt{s}$增长的过程都可以分为两个阶段：①线性增长阶段，即当 n_l 值较小时，n_l 随$\sqrt{s}$呈线性增长关系；②非线性增长阶段，当$\sqrt{s}$足够大，其相应的图像 Z' 和 Z''会无限接近原始信号 Z，因此，n_l 值会缓慢接近一个固定值。

然而，三类信号的 n_l 增长曲线出现拐点的位置（即由线性增长阶段转化为非线性增长阶段的位置）并不相同。这是由原始信号包含的细节所决定的，如果信号

波动较大(如 Type Ⅲ),则意味着图像可以包含更多的细节,需要较多的节点才能充分描述信号包含的细节。反之,若信号较为平滑(如 Type Ⅰ),则不需要太高的城市分辨率,即可反映出原始信号的基本状态。因此,随着城市像素数$\sqrt{s}$的增加,Type Ⅰ信号的城市分辨率会更早地进入非线性增长状态。

通过上述分析我们发现,在 n_l 与$\sqrt{s}$呈线性增长的阶段,其增长关系与原始信号的具体情况无关。因此,若在原始信号包含足够多细节的情况下,这种线性关系即可视为城市分辨率的变化规律。通过线性拟合,我们得到如下公式(3-8):

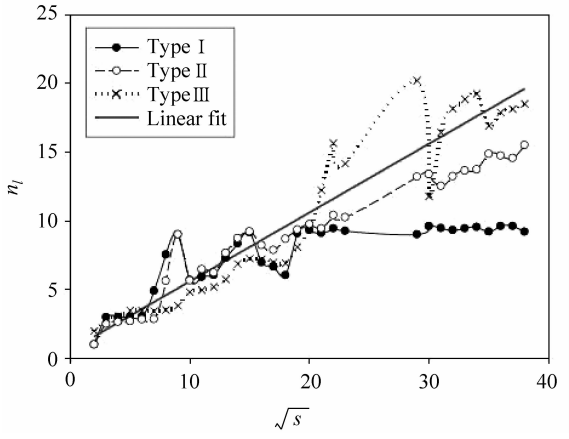
$$n_l = 0.9\sqrt{s} - 0.8 \tag{3-8}$$

至此,我们分析了原始信号的波动与网络规模对城市分辨率的影响,并得出了城市分辨率的线性变化规律。然而除了上述两个因素外,城市像素节点的分布情况对感知图像的影响也很大,现有的关于人/车辆轨迹分析的工作[11],已经总结出了如下人的移动特性:①异质的移动边界;②截断幂律分布的移动步长和等待时间;③截断幂律分布的相互接触时间;④分形的路径点。上述这些特性,使人群在任意时刻的分布情况都与符合均匀分布的随机节点差异很大,导致之前我们总结出的线性变化规律并不满足现实情况。为此,从实际应用的角度出发,我们进一步引入符合人的移动特性的城市像素节点,以分析城市分辨率的变化规律。

Self-similar Least-Action Walk(SLAW)[11]是一种随机游走模型,可以生成满足上述 4 点移动特性的轨迹数据。我们应用 SLAW 轨迹仿真数据,对图 3-9(a)(b)(c)中的三个信号进行采样,以模拟移动群智感知网络对环境现象的感知过程,并分别计算城市分辨率。图 3-11 列出了三个信号 n_l 随 $\sqrt{s}$的变化情况。与图 3-10 的结果类似,n_l 随 $\sqrt{s}$的变化情况同样分成了两个阶段,即线性/非线性增长阶段。对于图 3-11 中的线性增长阶段,我们进行了线性拟合,其公式如下:

$$n_l = 0.5\sqrt{s} + 0.6 \tag{3-9}$$

与公式(3-8)相比,公式(3-9)中拟合直线的斜率降低了,这意味着在相同网络规模的条件下,对同一信号进行感知,符合人的移动特性的网络,感知能力要差一些。其原因在于,人的移动特性使城市像素节点分布更不均匀。至此,我们通过仿真数据,得到了城市分辨率与城市像素之间的线性关系。



注：x 轴为$\sqrt{s}$，y 轴为 n_l，即$\sqrt{r}$

图 3-11　SLAW 轨迹的城市分辨率 r 与城市像素数 s 之间的关系

3.2.3　城市分辨率分析

城市分辨率是对感知图像的度量，而感知图像的质量又决定着移动群智感知网络对监测区域环境状态的捕捉能力，这种能力进而反映出移动群智感知网络的空间覆盖程度。考虑现实场景中环境、轨迹数据与仿真数据之间的差别，本节通过实测的环境与移动轨迹数据，对城市分辨率的变化规律作进一步分析。

我们分别运用了文献[24,25]中二氧化碳浓度和城市噪声观察数据作为原始信号 Z，并用北京、上海出租车轨迹数据[26]以及 KAIST 用户移动轨迹数据[27]作为城市像素节点的移动轨迹分别对环境信号进行采样分析。为了显示实际轨迹与仿真轨迹的一致性，我们同样利用 SLAW 轨迹进行了采样分析。

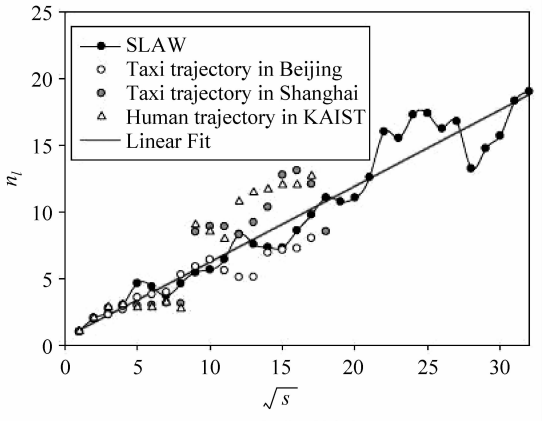
图 3-12(a)列出了二氧化碳浓度 n_l 与$\sqrt{s}$的计算结果。线性拟合公式如下：

$$n_l = 0.55\sqrt{s} + 0.56 \tag{3-10}$$

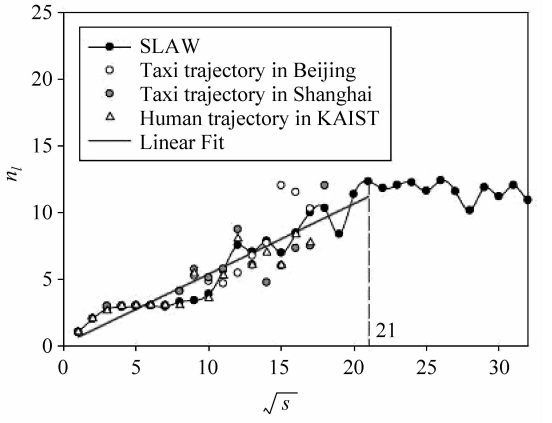
图 3-12(b)列出了城市噪声 n_l 与$\sqrt{s}$的计算结果。线性拟合公式如下：

$$n_l = 0.53\sqrt{s} + 0.10 \tag{3-11}$$

图 3-12(a)和(b)中拟合线的斜率与 3.2.2 节中仿真数据结果相符，进一步验证了城市分辨率与城市像素点之间存在着线性关系。得出这种线性关系的意义在于，在给定一个移动群智感知网络规模的情况下，可以根据这种线性关系，估算出该网络可以达到的城市分辨率(即空间覆盖质量)。另一方面，我们也可以反过来推出，需要部署多少移动感知节点能够达到所需要的城市分辨率需求。



(a) 二氧化碳浓度



(b) 噪声级别

图 3-12 城市分辨率 r 与城市像素数 s 之间的关系。城市像素的轨迹分别采用 SLAW 轨迹以及真实轨迹数据(北京、上海出租车轨迹,KAIST 用户移动轨迹)

3.3 地点覆盖率

上面提到的方法主要适用于城市环境监测、交通拥堵状况和道路健康状况监测等需要对整个城市的每个区域进行连续监测的大规模城市感知应用。与此不同,Chon 等人则研究了“以地点为中心”的移动群智感知应用的覆盖质量[28]。所谓“以地点为中心”的应用,就是自动识别或者跟踪用户每天访问的不同地点(如咖啡馆、超市、办公室、家、学校等),来帮助用户认识和分析自己的日常行为模式,或者获取基于位置的搜索和信息推荐等服务。构建和部署这些应用的前提是,对用户访问的每个地方采集足够的感知数据(如 GPS 位置、声音、图像、光照、Wi-Fi 信

号、指纹等)来建立各种模型。那么,这里的覆盖问题就是:在多长时间内有多少用户采集数据能够覆盖到多少人们经常访问的地点?为此,Chon 等人设计了以地点为中心的移动群智感知系统 CrowdSense@Place,他们收集了大量用户数据,从中得到了一些有趣的统计结果,并且分析了地点覆盖率与系统规模之间的关系。

3.3.1　CrowdSense@Place 系统及数据收集

CrowdSense@Place 系统由智能手机客户端 APP 和服务器基础设施两部分构成,如图 3-13 所示。用户可以使用智能手机 APP 采集图像、音频、Wi-Fi、GPS 等底层传感器数据,并且根据隐私设置有选择性地将数据上传到服务器,而服务器基础设施对这些感知数据进行存储,并利用光学字符识别(OCR)、对象识别、语音识别、声音分类等多种感知数据分类器提取高层语义信息,例如从图像中识别出汽车、建筑等对象,从音频中识别出说话内容,然后联合这些高层语音信息和一些底层感知数据推测出用户所在的地点。

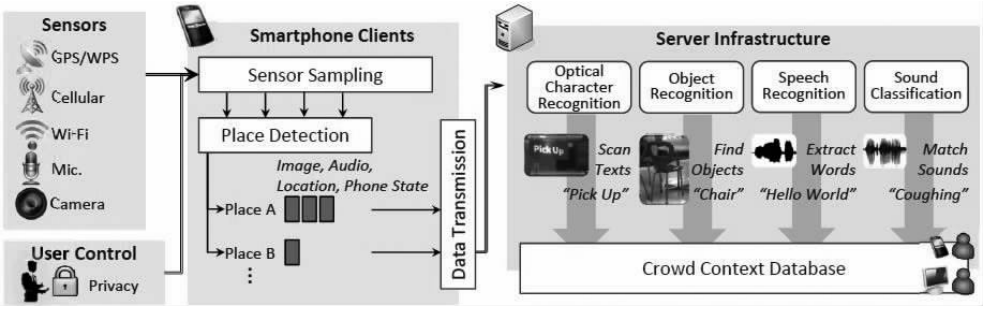


图 3-13　CrowdSense@Place 系统架构

Chon 等人在韩国首尔招募了 85 个实验用户,利用 CrowdSense@Place 系统在 2011 年 3 月至 2012 年 9 月期间收集了大约 4.8 万次用户访问不同地点的感知数据,数据量总计达 11GB,包含约 2.2 万个音频片段(190h 音频长度)和 6200 张照片,涉及约 1.35 万个地点。表 3-3 列出了该数据集的详细情况。

表 3-3　CrowdSense@Place 系统收集的数据集描述

感知数据类型		数　值
位置	覆盖面积	230.2km^2
	涉及的地点个数	13 447
	涉及的路径个数	483 379
	总访问次数	48 068
	数据量大小	813MB

续表

感知数据类型		数　值
照片	涉及的地点个数	1580
	照片数量	6242
	数据量大小	9.1GB
音频	涉及的地点个数	1517
	音频片段数量	22 604
	总长度	192h
	数据量大小	1.3GB

3.3.2　地点覆盖率分析

为了方便对地点覆盖率进行分析，Chon 等人从社交网络 FourSquare 中收集了同一时间段在韩国首尔产生的社交媒体数据作为参照，包括由 31 000 个用户产生的 1 078 100 个签到数据和 9200 张照片，涉及 98 899 个地点。统计结果发现，CrowdSense@Place 的 85 个实验用户访问了 FourSquare 所列地点中的 5999 个，约占总数的 6%。然而，对于某些类型的地点，实验用户有着更高的覆盖率，例如，大学的覆盖率为 14%，艺术与娱乐中心的覆盖率为 9%。同时，地点的流行度越高，覆盖率也越高。图 3-14 将 FourSquare 中的地点按照签到数量确定流行度，并显示了每个流行度级别的实验用户覆盖率。例如，85 个实验用户对流行度前 0.4%的地点（签到数量大于 1000）的覆盖率为 22%。这说明，仅仅利用少量的用户，就能对人们常去的地点提供相对较高的覆盖率。

3.3.3　地点覆盖率预测

除了对现有实验数据分析之外，我们还希望通过对少量用户采样就能够刻画出地点覆盖率与用户数量之间的一般关系，从而在需要更高覆盖率的时候，能够估算出还需要多少用户。这种预测是完全可行的，因为随着流行度的增加，新的用户更可能与现有用户具有相似的访问地点和路径偏好。为此，Chon 等人提出了一个数据驱动的估计模型来预测地点覆盖率与用户数量之间的关系。该模型主要基于两个基本的观察结果：首先，用户对地点的访问次数服从幂律分布。如图 3-15(a)所示，对 FourSquare 中所有地点的签到次数的累积分布函数分别使用双曲线衰减函数和帕累托分布进行拟合，可获得较高的拟合度（决定系数分别为 0.97 和 0.61）。该结果表明许多处于长尾分布的“长尾”位置的地点将很难被覆盖到。其次，一个用户对某个地点的访问概率随着该地点与用户最常访问地点的距离增加

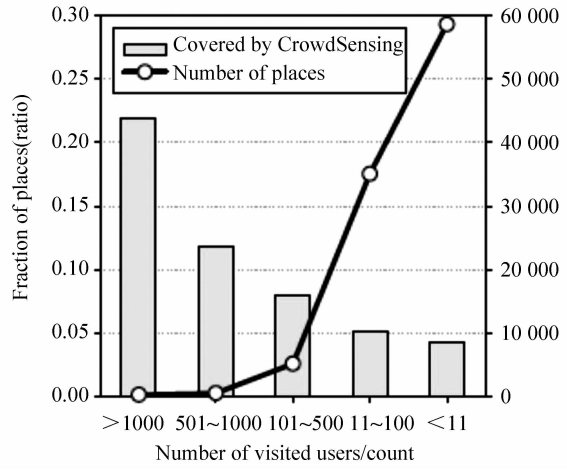
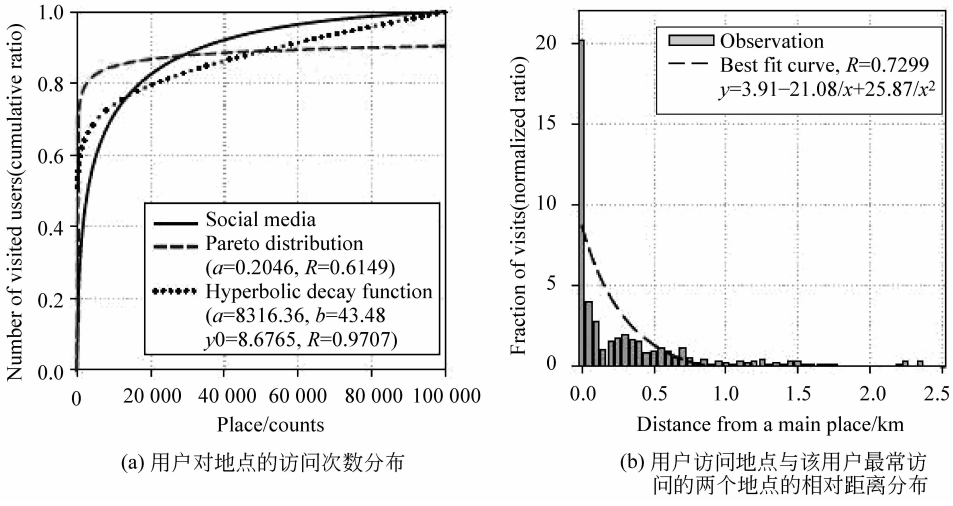


图 3-14　不同流行度地点的覆盖率



(a) 用户对地点的访问次数分布

(b) 用户访问地点与该用户最常访问的两个地点的相对距离分布

图 3-15　用户对地点的访问服从幂律分布

而急剧减小。图 3-15(b)显示了用户访问地点与该用户最常访问的两个地点的相对距离分布直方图，该分布可以使用一个二次反比函数进行拟合(决定系数为 0.72)。该结果表明，用户访问的地点主要集中在少数几个用户经常访问的地点，例如用户的居住地和工作地。根据这两个分布规律，可以通过仿真的方式产生新的用户，从而生成一个用户模型。为了更加准确，Chon 等人将地点按照 FourSquare 数据集中的流行度分为 4 类(签到数量分别为 1000＋、100～999、10～99、1～9)，并且分别估计模型中两个分布函数的参数。图 3-16 表示预测模型中地

点覆盖率与用户数量之间的关系。在用户数量较少的情况下，可以与真实情况进行对比，结果显示预测误差只有 11.3%。同时，该模型预测出，还需要大约 2.5 万个用户才能实现在两个月的时间段内对韩国首尔地区流行地点（访问次数大于 100）的覆盖率达到 60%。这个人数需求仅占韩国首尔总人数的 0.2%，比较容易满足。例如，一个用于记录首尔地区公交车到达信息的手机应用 SeoulBus 就有超过 26.8 万的用户数量。

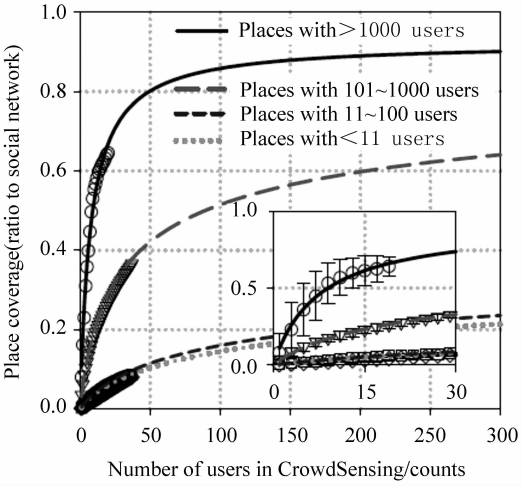


图 3-16 预测模型中地点覆盖率与用户数量之间的关系

3.4 基于压缩感知的感知质量增强

压缩感知是近年来兴起的一种信号重建技术。该技术的创新之处在于它能够实现在远低于香农-奈奎斯特采样率的情况下，对信号进行高质量的重建。压缩感知技术对信号的重建能力，恰恰满足移动群智感知网络中对“感知盲区”实现数据填充的需求。以城市环境温度监测为例，我们可以利用移动群智感知网络收集一部分区域的温度数据，这实际上是对城市温度信号的一次有偏采样，之后再借助压缩感知技术重建城市温度信号。但是，实现压缩感知技术需要满足一个理论前提：待感知信号应能在某个域（基）上被稀疏表达（即待感知信号应具有隐含结构）。

文献[29]对几组公开的环境监测数据的隐含结构进行了分析。这些数据包括：在室内[30]、森林[31]以及海洋[32]环境下采集的温度和光照强度数据。首先，文献[29]以环境矩阵的形式对这些数据进行建模。这里环境矩阵 X 被定义为

$$X = (x(i,j))_{n\times t} \tag{3-12}$$

式(3-12)中，n 表示数据采集位置，t 表示数据采集时间。如图 3-17 所示，文献[29]对环境矩阵的数据进行奇异值分解，可以看出环境矩阵的奇异值衰减的速度非常快，这意味着环境矩阵具有低秩性，即含有大量冗余信息。此外，这种现象从一个侧面表明环境监测数据能被稀疏表达。因此，当仅获得少量环境感知数据的情况下，可以通过压缩感知技术，对环境矩阵进行恢复。

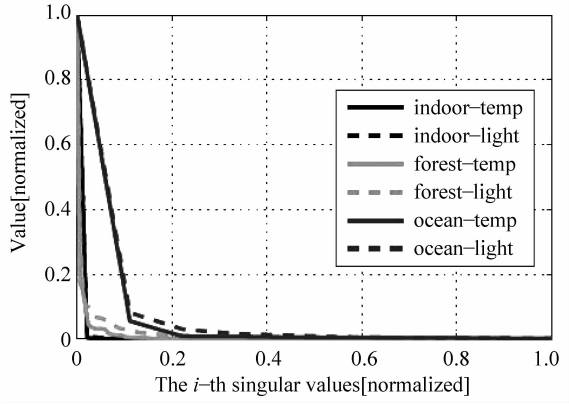


图 3-17　环境矩阵低秩性分析

这里，压缩感知的过程可以形式化为如下优化过程：

$$\begin{aligned} &\text{优化目标：}\min(\operatorname{rank}(\hat{X})) \\ &\text{约束条件：}B \cdot \hat{X} = X \end{aligned} \tag{3-13}$$

式(3-13)中，$\hat{X}_{n\times t}$表示感知矩阵，即由采集到的感知数据构成的矩阵；B 表示标识矩阵，即用来表示感知矩阵在相应位置上是否含有环境质量的感知数据，具体定义如下：

$$B = (b(i,j))_{n\times t} = \begin{cases} 0, & \text{如果 } \hat{X}(i,j) \text{ 数据丢失} \\ 1, & \text{否则} \end{cases} \tag{3-14}$$

通过优化目标函数，即可实现对感知数据的恢复。

3.5　基于多源数据相关性的感知质量增强

在大数据时代，丰富的数据收集手段与数据集使我们有充足的数据源去挖掘不同类别数据之间的相关性。例如，文献[25]表明，城市空气污染物分布情况与城市 POI(兴趣点)数据、交通流量数据密切相关。如果我们有城市区域 A 的空气污染指数，而没有区域 B 的空气污染指数，但我们通过分析发现，区域 A 与 B 的 POI 分布、交通流量状况类似，由此我们可以推断区域 A 与 B 的空气污染指数类似。