

西班牙三大移动网络运营商数据采集辅助下的下一代网络规划转型

M. Umar Khan¹

Abstract— 在本文中,我们探讨了提取有关 5G 网络规划所需有用信息并进行网络规模设计的必要性。这些数据可以帮助我们实现更高效的网络规划设计,尤其是在拓扑结构和成本方面。为此,使用了西班牙三大移动网络运营商 (MNOs) 的真实开放数据库来估算流量,并识别用户密度最高的区域以部署新服务。我们提出了用于清理数据库、提取有意义的交通信息以及为新的 gNB 部署可视化流量密度模式的数据采集程序。我们介绍了网络数据领域的最新进展。详细描述了所考虑的网络数据库。解释了网络数据采集实体及其提出的程序。讨论了相应的结果,并得出了结论。

I. 介绍

蜂窝网络数据在早期的研究中就受到了关注,用于分析订户 [1] 的时空变化情况,识别网络流量模式 [2],并提升最终用户的 QoE[3]。对网络数据的研究具有多学科性质,涵盖了与信号处理、移动通信和数据挖掘参数 [4] 相关的交通管理方面、服务提供以及城市动态。网络数据的可用性可以为实现增强性能带来显著影响,并对网络设计、成本和拓扑结构 [5] 产生深远影响。然而,需要深入的知识来理解网络数据来源、可访问性、特性和应用,以提取特定用例相关的信息 [6]。此外,这些研究是在从不同来源获得的数据上进行的,例如移动电话、部署的传感器、云无线接入网 (C-RAN) 以及相应 MNO 自行收集的数据。由 MNO 拥有的数据或从 C-RAN 获得的数据是私有的,因为它们包含与运营商相关的敏感信息,可能不会公开提供。然而,由于存在可公开访问的众包数据库,研究人员可以接触到能够分析和比较不同网络及运营商性能的数据,而不必依赖于 MNO 拥有的数据。例如,在 [7] 中,作者利用一个真实世界的众包数据库来分析传统的 LTE 部署及其承载的流量。所提出的方法旨在识

别现有的部署情况,目标是处理当前和未来的交通需求以及提高网络容量的可能性。结果显示,已部署的 LTE 小区相较于当前的交通需求而言配置过多,尽管仍需显著增强容量以服务于未来负载。[8] 的工作提出了一种基于众包数据库的方法论,用以推断现有基础设施的状态,并分析多个移动网络运营商的网络密度和频谱资源。结果显示,在密集的城市区域,LTE 基站的部署是以容量为导向来应对更高的流量需求。另一方面,郊区部署的资源是通过部署较低频率的频谱以降低成本而覆盖范围为导向的。从所提出的方法论中获得的科研成果被讨论应用于新 5G 部署的设计。

在此背景下,众包指的是从社区或特定工作组获取输入。更具体地说,数据众包涉及数据采集、管理和分析以实现针对某一特定问题的适当且有意义的解决方案 [9], [10]。使用开放的众包数据已经拓宽了研究领域的范围,涵盖了从规划智慧城市到业务发展或数据分析 [11] 等各个方面。在无线通信领域,众多研究已经通过利用众包数据库展示了重要的发现。例如, [12], [13] 的作者提出了一种数据采集和分析模型来衡量 LTE 网络面对不同威胁的安全水平。在 [14] 中,作者倾向于使用众包数据库对 LTE 小区中的电磁场 (EMF) 暴露进行评估和持续监测,而不是传统方法。他们表明可以从众包获得的信号强度测量值提供合适的 EMF 强度估计。一项基于移动应用的真实世界众包数据库的研究识别了交通 [15] 的多个方面,例如车辆对全球交通的影响以及从交通数据中获取的不同架构性能的有用见解。[16] 的作者提出了一种模型,分析与时间、信号强度、信号质量和 MNO 等其他网络变量相关的 LTE 下行链路吞吐量。通过数据分析他们得出结论认为,相比于信号强度,信号质量是预测 LTE 用户吞吐量的适当变量。

同时,使用众包数据的好处是可以利用一系列参数和指标来推导出有意义的信息。例如演化的节点 B

*This work was not supported by any organization

¹M. Umar Khan is with Faculty of Computer Engineering, COMSATS University Islamabad, Pakistan
umar_khan@comsats.edu.pk, www.drmukhan.com

(eNB) 密度、参考信号接收功率 (RSRP)、参考信号接收质量 (RSRQ) 以及可用带宽都被审查以观察在不同地理区域的 LTE 网络中频谱是如何被使用的 [17]。在 [18] 中, 作者研究了遗留基础设施以推断运营商之前的 LTE 部署策略是否是出于覆盖范围或容量考虑。分析显示, 大多数运营商的策略旨在提供目标区域内的覆盖范围而非容量增强。

鉴于之前的讨论, 我们观察到在网络规划阶段可以考虑网络数据的几个方面以获得重要的见解。在这方面, 我们注意到从公开数据库获取的数据可用于识别交通密度最高的区域, 这对于网络容量设计是一个相关的输入。这同时也带来了在网络安全成本方面的额外好处。在这项研究中, 使用了 OpenCellID 数据库中的网络数据来获取关于与交通密度相关的运营商遗留网络状态的信息。选择 OpenCellID 数据库是因为现有蜂窝网络的数百万样本对研究界公开可用, 并且这些样本将 LTE 标识符与交通密度相关联。因此, 在本文中我们重点关注利用 OpenCellID, 其详细信息在 §II 中提供。

A. 贡献

本文提出了一种数据采集方法, 该方法能够为新 5G 服务划定目标部署区域。简言之, 本文做出了以下贡献:

- 1) 我们提出利用实际网络数据来估计网络流量及用户集中度最高的目标部署区域, 反映出最拥挤的区域。
- 2) 我们开发并提出了一种数据采集程序。该提出的程序利用标准的 LTE 标识符来确定流量最高的区域。
- 3) 我们已经对来自西班牙三大移动网络运营商的数据分析程序进行了评估。

以下各节描述了网络数据采集的完整过程。我们首先详细描述众包数据库, 然后讨论该流程。

II. OPENCELLID 数据库

OpenCellID 是一个开放访问的数据库/平台, 移动设备通过众包应用程序 [19] 向其发送与连接相关 (例如位置 and 不同标识符) 的数据。该应用程序记录了设备和基站 (BS) 的全球定位系统 (GPS) 坐标 (纬度、经度)、无线接入技术 (RAT)、移动国家代码 (MCC)、移动网络代码 (MNC)、跟踪区域码 (TAC)、小区塔

ID (CID)、小区可能存在的近似范围 (范围, 以米为单位), 以及记录的样本数量。当设备与基站建立连接时, 它会扫描上述变量的所有信息, 并将每次扫描, 表示为样本, 由设备记录并提交到数据库。该数据库还提供了有关 BS 位置的信息, 这些信息可能是从样本中获得的, 也可能是移动网络运营商 (MNO) 本身提供给数据库的。

A. 数据库矩阵

OpenCellID 数据库包含需要处理才能正确使用的原始数据。我们考虑西班牙的原始数据, 以便了解根据 5G 国家计划 [20] 部署新技术所需的网络基础设施。首先, 通过从原始数据库中提取对应于西班牙 MCC=214 的信息来形成网络数据库矩阵 $\mathbf{D}$ 。在矩阵 $\mathbf{D}$ 中, E 是样本的数量 (以百万计), 并且每行 $\mathbf{d}_e$ 的 $\mathbf{D}$ 表示一个数据点 $\mathbf{d}_e = (r_e, m_e, t_e, n_e, l_e, s_e)$ 。因此, $\mathbf{D} := \{\mathbf{d}_e\}_{e=1}^E \forall \mathbf{d}_e \in \mathbb{R}^v$, 其中 $v = 6$ 代表数据变量的数量。

$$\mathbf{D} = \begin{pmatrix} r_1 & m_1 & t_1 & n_1 & l_1 & s_1 \\ r_2 & m_2 & t_2 & n_2 & l_2 & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_e & m_e & t_e & n_e & l_e & s_e \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_E & m_E & t_E & n_E & l_E & s_E \end{pmatrix},$$

并且数据点 $\mathbf{d}_e$ 的分量是:

- r_e : 行 e 的无线接入技术 (RAT)。
- m_e : 第 e 行的移动网络代码 (MNC)。
- t_e : 第 e 行的跟踪区域代码 (TAC)。
- n_e : 第 e 行的小区 ID (CID)。
- l_e : 具有纬度和经度的二维单元位置 $l_e = (l_e^1, l_e^2)$ 。
- s_e : e 的样本数量。

数据集包含西班牙现有 GSM、UMTS 和 LTE 小区的坐标。在本研究中, 我们将遗留 LTE 小区视为估算蜂窝网络当前数据使用情况的数据源, 因为 GSM 和 UMTS 对当今流量贡献甚微。该数据集包含了数百万个 CIDs (小区标识), 这足以进行网络分析以推断基础设施和交通信息。然而, 数据库并未提供 eNB-IDs (演进节点 B 标识) 以保护移动网络运营商的隐私。另一方面, 访问 eNB-IDs 可以简化数据获取过程, 并使流量信息能够迅速用于分析。在缺乏 eNB-IDs 的情况

下，则需要一个从唯一识别的小区中提取交通信息的数据采集程序。

III. 数据采集程序

在上一节中，我们介绍了如何从 OpenCellID 数据库形成行数据矩阵 $\mathbf{D}$ 。然而，这些信息必须经过处理以服务于特定目标。一个可能的目标是识别具有最大用户密度的 5G 部署区域 (5GDA)，这将代表交通量最高的地区。本节介绍图 1 中的网络数据获取过程。该过程的核心在于确定包含样本数量最多的跟踪区，这个跟踪区被标识为交通量最高的 TAC。另一方面，我们也需要保证数据集的可靠性，例如避免在选定基站中出现重复。为此，在 §III-E 中设计了网络数据钻探 (NDD) 过程。该过程包括两个阶段，预选和后选，分别描述在 III-C 和 III-D 中。

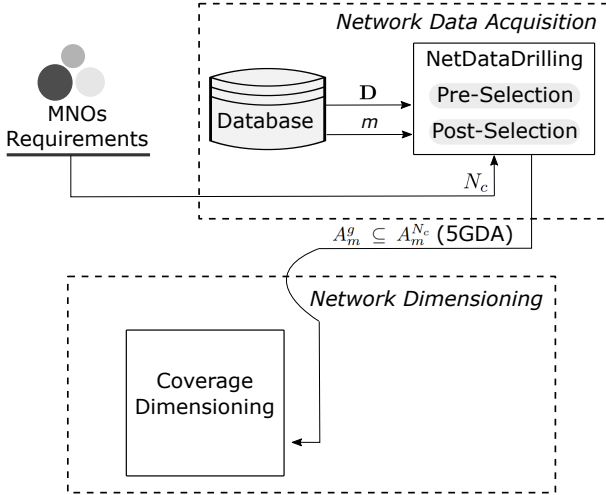


Fig. 1: 网络数据采集实体与网络规划相关联，以提供作为覆盖规划输入的 5GDA (A_m^g)

A. 最高流量 TAC 区域

最高流量 TAC (HTTAC) 区域由网络数据库 $\mathbf{D}$ 提供的样本数量 s_e 最多的那些小区组成。在 $\mathbf{D}$ 中，每个 MNO_m 都有一个 TAC 集合 $\mathcal{T}_m = \{t_1, \dots, t_j, \dots, t_J\}$ ，其中 t_j 代表一个单独的 TAC；一个单元格集合 $\mathcal{N}_m = \{n_1, \dots, n_k, \dots, n_K\}$ ，其中 n_k 表示 CID 和 $N_m = |\mathcal{N}_m|$ 是 MNO_m 的总单元格数；以及 $s_m^{j,k}$ 表示来自单元格 k 内的第 j 个 TAC 中的样本数量。¹ 我们假设每个小区的最大样本数代表该区域最高的交通密度。因此，从

¹ 请注意，如果单元格不属于跟踪区域，则为 $s_m^{j,k} = 0$ 。

跟踪区 TAC 内所有现有小区的总样本数最大值中获得最高交通 TAC 或 HTTAC 的 ID 为: H_m

$$H_m = \arg_j \max H_m^j, \quad (1)$$

$$H_m^j = \sum_{k=1}^K s_m^{j,k}, \forall j = 1, \dots, J.$$

B. 5G 部署区域 (5GDA)

部署区域是指为安装新的 BS 以提供新服务而划分的地理区域。由于 5G 网络必须与 LTE 网络相比，每比特成本更低地提供新服务，更多的订阅数量会增加网络利用率，并且同时帮助 MNO 降低每比特成本。此外，每个 MNO 可能有自己的关于 5G 基础设施部署成本的推出策略。因此，有必要识别出交通量最高的区域。我们继续选择 N_c 个传统小区在 H_m 中，使得 $N_c \subseteq \mathcal{N}_m$ ，其中 N_c 是那些支持最高流量密度的小区。这些小区提供了一个估计，即大多数订阅者所在的交通量最高的区域，这样 MNO 可以开始部署新的 5G 服务以尽量减少成本。覆盖 N_c 个小区的地理区域由 $A_m^{N_c}$ 表示。此外，MNOs 可能在该地区部署 5G 服务时可能会受到财务限制，这通过 N_c 单元的价值来解决。如果选择较大的 N_c 值，则区域 $A_m^{N_c}$ 将更大，并且 MNO 需要在更广泛的区域内部署其服务，导致成本更高并无法满足财务预算。因此， N_c 的值转化为 MNO 的财务和覆盖区域限制。此外，我们将 A_m^g 定义为高流量集中区，该区域通过数据可视化确定，在此将部署 gNB 以提供 5G 服务，使得 $A_m^g \subseteq A_m^{N_c}$ 。

C. 预选

预选阶段是 NDD 程序的第一阶段（参见图 1），从全球数据库 $\mathbf{D}$ 中提取特定国家 (MCC) 及其网络 (MNC) 的信息。然后，我们识别出主要的移动网络运营商，并通过与外部数据源进行比较来验证数据库中提供的 RAT 是否正确，如下所述：

- 首先，我们提取属于西班牙的网络数据子集（国家代码 MCC=214）。
- 我们根据各自在西班牙的市场份额选出前三大 MNOs [21]。
- 我们确认由 OpenCellID 数据库提供的 LTE RAT 已经被运营商实际部署，通过将此信息与 [22] 进行对比。

请注意,通过简单选择不同的 MCC,可以处理来自任何其他国家的数据。类似地,顶级移动网络运营商的数量也可以自定义。

D. 选择后

后选择阶段是 NDD 程序的核心,确保分析的质量,避免数据冗余,并揭示模式以识别订户密度较高的区域。为了保证分析质量,我们将置信基准样本(CBS)区间定义为传统小区样本数在 $[a, b]$ 范围内的数量,从而保证 N_c 小区的存在。请注意,在数据库中存在记录样本数量较少的小区(例如低于 100),这种情况可能代表了后续阶段安装、移除、重新配置或合并过的小区。也有一些小区的样本数与大多数小区相比非常大(超过 2000),被视为异常值。因此,我们通过定义 a 和 b 分别为下限和上限,不将这些小区纳入我们的分析。

接下来,我们将两个 LTE 标识符组合以唯一确定小区,称为 TAC 和 CID 组合。一旦小区被唯一识别,我们获得流量最高的区域。从 HTTAC 区域中,我们识别出 MNOs 选择一个可行子区域的小区,即 5GDA。5GDA 作为输入提供给网络定型实体(参见图 1)用于覆盖定型。

后选择阶段考虑了以下方面:

- 数据库中存在单个小区的多个样本。一个 CID 可能对应多个记录的 GPS 坐标。为了避免同一真实位置处多个小区的冗余,我们定义 TAC 和 CID 的组合来识别单独的小区。
- 为了估计每个 MNO 的实际交通量,我们假设样本数量最多的区域(TAC)在该区域内支持最高的交通密度,称为 HTTAC。
- 我们然后选择 HTTAC 内的 N_c 单元格,使得这些单元格具有最多的样本数,即 UE 建立的数量。
- 最后,选择了子区域 $5GDA A_m^g \subseteq A_m^{N_c}$,代表了最高的人流量密度区域。

E. NetDataDrilling 算法

本小节描述了数据选择和提取的过程网络数据钻探(NDD),并提供了算法 1 的伪代码。它需要以下输入:

- 网络数据库 D 。
- 表示由 MCC 代表的国家代码 M 。
- 用于选择旧单元的常数 N_c 。

- 主要三家运营商用移动网络代码 m 表示。
- 常数 a 和 b 分别作为 CBS 方法的下界和上界。

NDD 程序的结果是 H_m , 每个 MNO 的 HTTAC ID 为 m , 以及具有 N_c CIDs 的小区集合 $\mathcal{N}_c$ 。

在第 5 步中,预选过程提取了与国家代码 M 对应的数据库以及 m 个 MNO。在这个过程中,针对每个 MNO(步骤 6 中的对于循环),一些辅助变量存储中间结果。 S_{aux} 表示每个 TAC 内那些小区样本的总和,因此其维度等于跟踪区域的数量。 V_{aux} 表示 HTTAC 内部每个小区样本的总和(维度等于构成 HTTAC 的小区数量),这些小区的 CIDs 存储在 W_{aux} 中。

我们首先移除数据点 d_e , 使得它们的样本数量超出 CBS 界限,即: $a > s_e > b$ (步骤 7-11)。接下来,我们找到唯一的 TAC 和 CIDs。由于相同的单元格(相同的 CID)可能存在多个版本,因为记录的 GPS 坐标不同,因此 TAC 和 CID 的组合确定了唯一的 TACs 和 CIDs,并消除了冗余。过程 UNIQUE 使用对 (TAC, CID) 来识别唯一性并将 TACs 和 CIDs 的索引分别存储在 $\mathcal{T}_m$ 和 $\mathcal{N}_m$ 中(步骤 12)。

在步骤 13-20 中,每个 TAC 的样本总数存储在 S_{aux} (第 17 步)中,通过将每个 TAC 关联的所有 CID 的样本数相加得到。我们在第 14 步开始循环,在 k 中保持 $\mathcal{T}_m$ 中的 TAC 索引(由 j 给出),对应于该迭代 e 的循环中的 TAC (t_e)。对于 $\mathcal{T}_m$ 中的每个 TAC 编号,我们将它们的编号与数据库 $D(t_e)$ 中的进行匹配,并检查所分析行的单元格 e (用 $D(n_e)$ 表示)是否属于 MNO 的网络(步骤 16)。这里, $D(t_e)$ 和 $D(n_e)$ 分别代表来自 D 的 TAC ID (3rd 列) 和 CID (4th 列) 在行 e 中。在步骤 21 中,我们找到 MNO m 的流量最高的跟踪区域 H_m , 对应记录样本数量最多的 TAC。

在步骤 (22-26) 中,我们计算位于目标区域内的单元格的样本总数 H_m 。因此,一旦找到 H_m , 在 D 中的对应于位于 H_m 内单元格的样本数量被存储在 V_{aux} 中(步骤 24)。然后将位于目标区域内的单元格的 CIDs H_m 存储在 W_{aux} 中(步骤 25)。最后,我们需要识别支持最高流量的 N_c 单元。在第 27 步中,与 H_m 相对应的单元按照样本数量降序排列。然后,如步骤 28 所示,选定的 N_c 个小区是集合 $\mathcal{N}_c$ 中的前 N_c 个位置,其中 $n_{c,k}$ 表示在 H_m 内流量第 k 高的小区的 CID。

1) NDD 算法的复杂度: 算法 1 的复杂性基于数据库 D 的大小,该数据库由样本数量 E 表示。为了处理

Algorithm 1: 网络数据钻探

```

1 procedure
  NetDataDrilling( $\mathbf{D}, M, N_c, m, a, b$ )
2  $S_{aux} \leftarrow 0$ 
3  $V_{aux} \leftarrow 0$ 
4  $W_{aux} \leftarrow 0$ 
5  $[m, \mathbf{D}] \leftarrow \text{Preselection}(M, \mathbf{D})$ 
6 for all  $m$  do
7   for all  $e$  do
8     if  $(s_e < a \vee s_e > b)$  then
9        $\mathbf{D} \leftarrow \mathbf{D} \setminus d_e$ 
10    end
11  end
12   $[\mathcal{T}_m, \mathcal{N}_m] \leftarrow \text{UNIQUE}(\mathbf{D}(t_e, n_e)) \forall e$ 
13  for all  $t_e \in \mathcal{T}_m$ 
14     $k \leftarrow j : t_e = \mathcal{T}_m(j)$  do
15      for all  $e$  do
16        if  $t_e = \mathbf{D}(t_e)$  and  $\mathbf{D}(n_e) \in \mathcal{N}_m$ 
17          then
18             $S_{aux}(k) \leftarrow S_{aux}(k) + s_e$ 
19          end
20        end
21       $H_m = \max S_{aux}$ 
22      for all  $e : \mathbf{D}(t_e) = H_m$  do
23        Find index  $k : \mathcal{N}_m(k) = \mathbf{D}(n_e)$ 
24         $V_{aux}(k) \leftarrow V_{aux}(k) + s_e$ 
25         $W_{aux}(k) \leftarrow \mathcal{N}_m(k)$ 
26      end
27       $\mathcal{N}_c \leftarrow \text{SORT}(V_{aux}, W_{aux})$ 
28       $\mathcal{N}_c \leftarrow \{n_{c,1}, \dots, n_{c,k}, \dots, n_{c,N_c}\}$ 
29 end

```

$\mathbf{D}$, 需要一个有限次数的迭代, 这由对于循环给出。因此, 由内部循环表示的算法 1 的复杂度可以计算为:

$$\mathcal{O}(E) + \mathcal{O}(E \times |\mathcal{T}_m|) + \mathcal{O}(|\mathbf{D}(t_e)|). \quad (2)$$

请注意, (2) 提供了每个 MNO_m 所需的复杂度, 在实践中, 由于可以从每个 MNO 独立并行处理样本, 这将是总复杂度。

IV. 结果

第一步是创建由 NDD 算法使用的数据库 $\mathbf{D}$, 该数据库是从 OpenCellID 数据生成的。对应西班牙的累积样本分别为 GSM + UMTS + LTE, 记录每个 MNO 的数据为 270 万、260 万和 310 万 (参见图 2), 分别对应 MNO1、MNO2 和 MNO3, 这些数据对于我们的分析来说是足够的。我们在地图上绘制了三个 MNO 对应的样本, 如图 3(a,c,e) 所示, 通过绘制小区基站来实现。

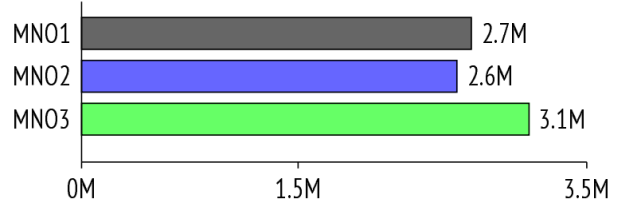
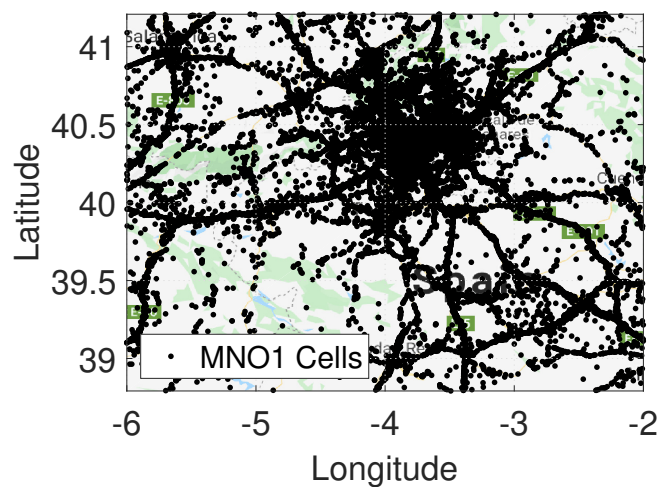


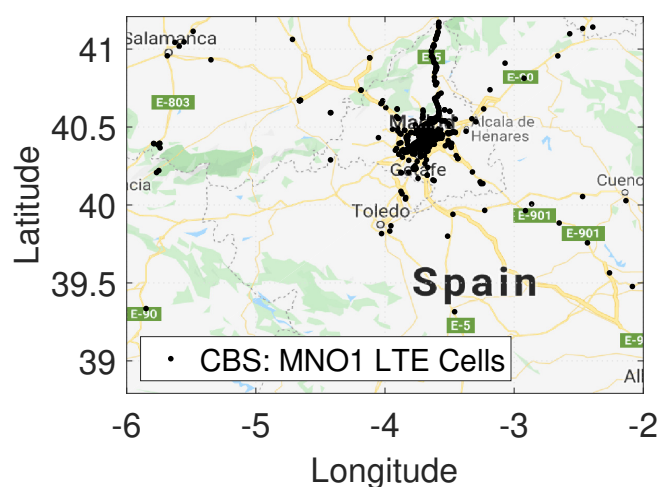
Fig. 2: 前三大西班牙移动网络运营商的累计样本 (GSM + UMTS + LTE)

NDD 过程在算法 1 中通过设置 $a = 100$ 和 $b = 1000$ 采用 CBS 方法。满足 CBS 标准的 LTE 小区如图 3(b,d,f) 所示。这些小区以千为单位 (3938, 3404, 3665) 分别对应 MNO1、MNO2 和 MNO3。现在确定唯一的 TACs 和 CIDs。对图 3(b,d,f) 所绘制区域的单个 LTE 小区进行检查, 结合唯一的 TAC ($\mathcal{T}_m$) 和 CIDs ($\mathcal{N}_m$), 以获得唯一 LTE 小区的数量。组合的输出显示 MNO1 有 3895 个, 而 MNO2 和 MNO3 分别有 3296 和 3408 个单元格, 市场份额分别为 $\kappa_1 = 23.55\%$ 、 $\kappa_2 = 25.79\%$ 和 $\kappa_3 = 30.14\%$, 分别是 [21]。

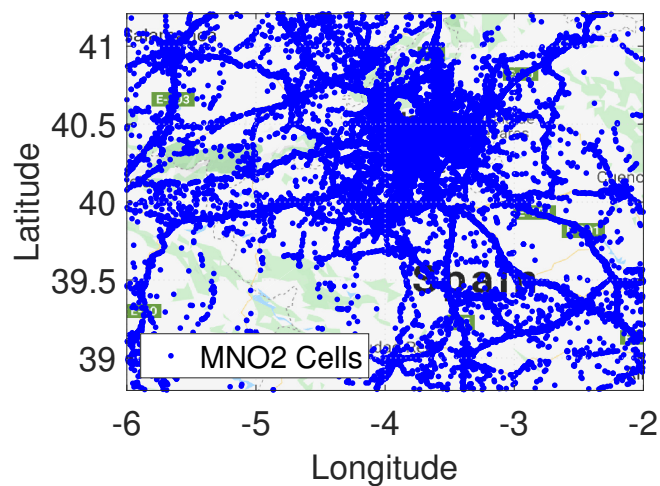
最后, 确定了 5G 部署区域选择 (A_m^g)。网络数据分析指导 MNO 在规划过程中选择那些高度集中的区域, 即 5GDAs, 用于新的部署。三个 MNO 的 HTTACs H_1, H_2 和 H_3 如图 4(a,b,c) 所示。我们发现 H_1 和 H_2 所覆盖的区域都位于马德里的中心, 而 H_3 则位于巴塞罗那的中心 (见图 4(c))。为了定位 5GDA, 我们设置了 $N_c = 100$ 以找到每个 H_m 内的流量最高的小区, 如 MNO1 所示的图 4(b)。MNO1 的 HTTAC 呈现出头/尾模式, 在这种模式中, 高密度的小区代表 HTTAC 的头部, 而尾部则表示较低浓度的小区。同样的模式在图 4 (c 和 f) 中也可以观察到, 其中尾部位于 HTTAC 的上部, 头部则位于底部。然而, 在 MNO2 的情况 (图 4, b 和 e) 中, 模式是均匀的, 细胞集中在城市的中心, 而沿着道路两侧的数量较少。



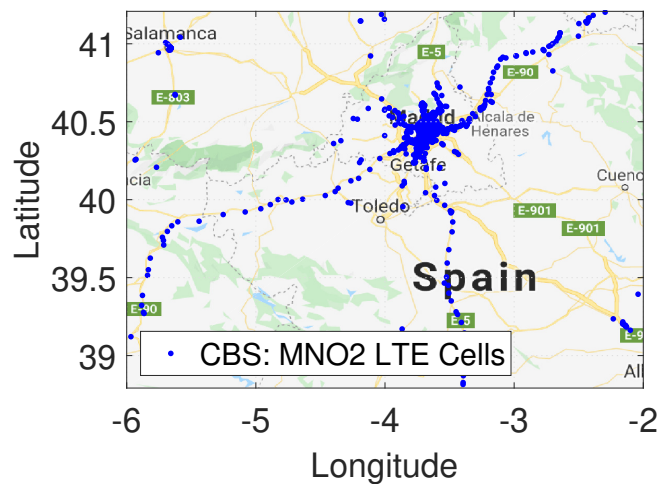
(a)



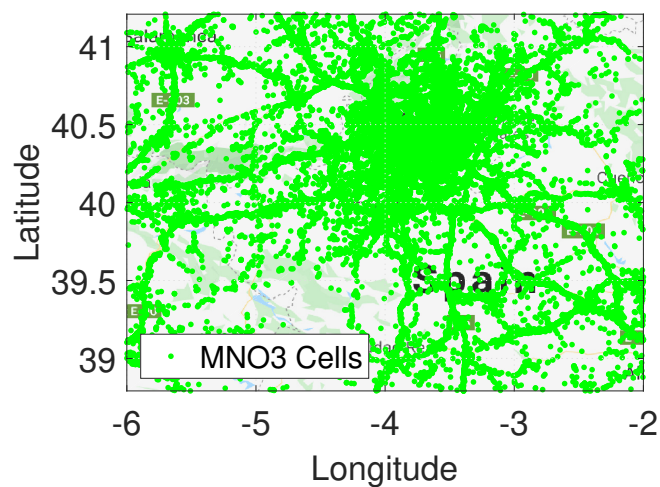
(d)



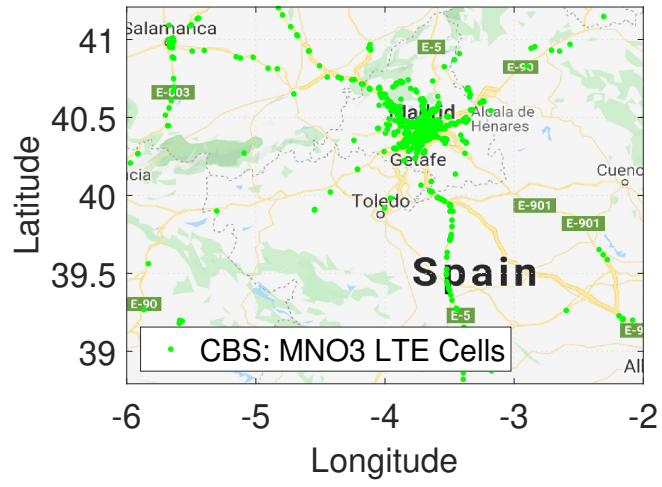
(b)



(e)

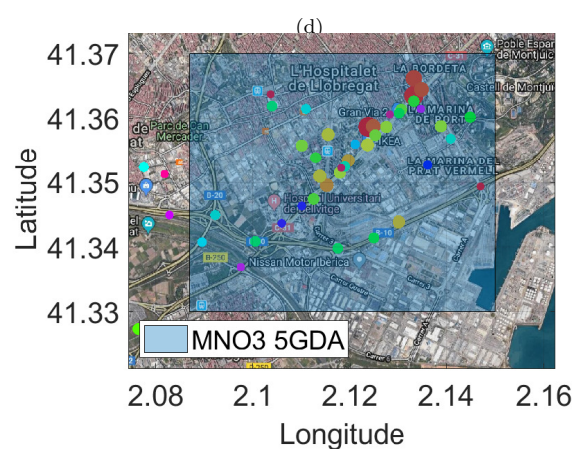
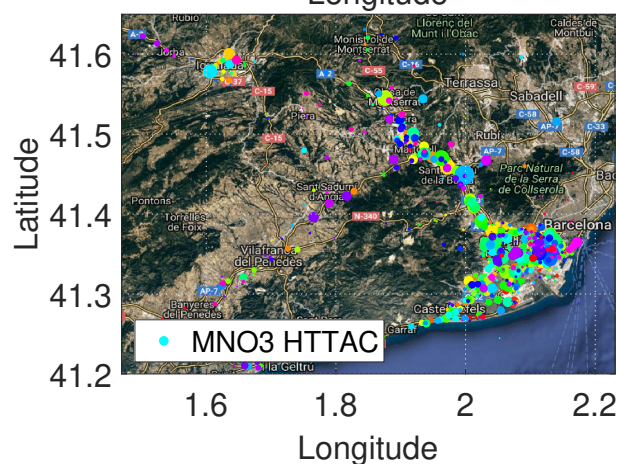
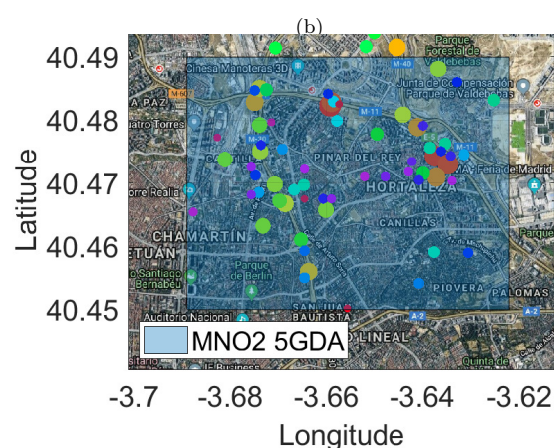
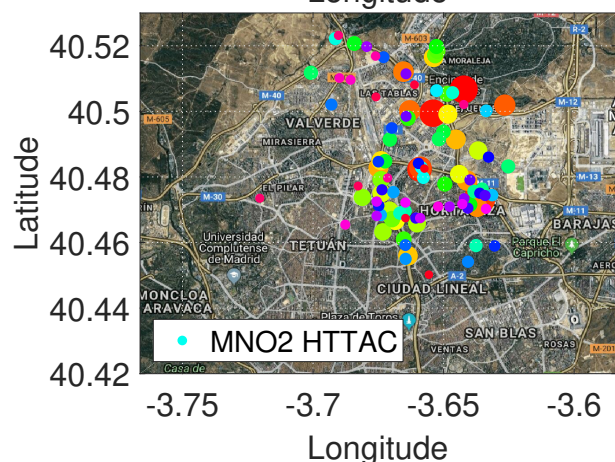
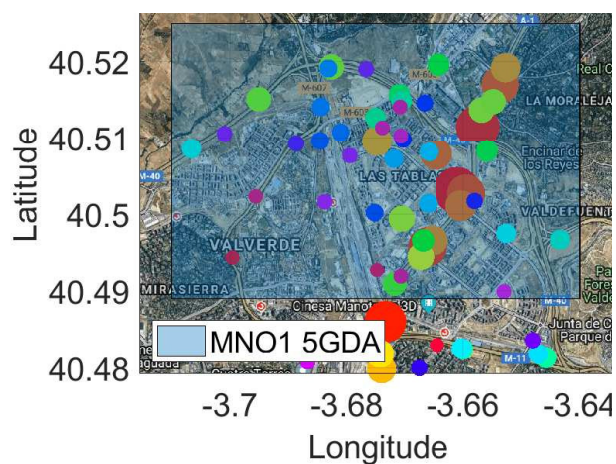
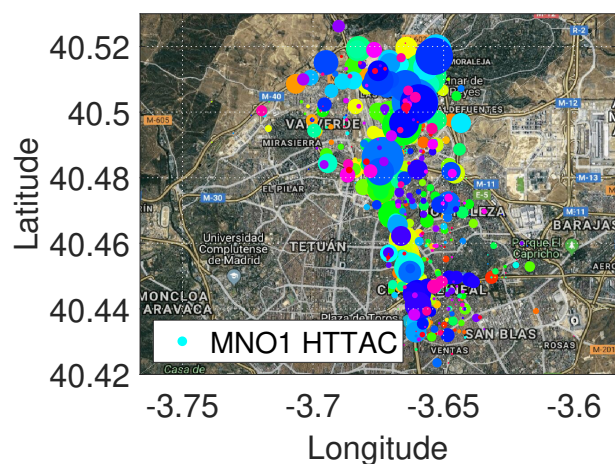


(c)



(f)

Fig. 3: 累积样本 (GSM + UMTS + LTE) (a,c,e) 和作为 CBS 结果的 LTE 样本 (三个西班牙移动网络运营商 b,d,f)



(e)

(f)

Fig. 4: 彩色圆圈代表流量最大的 LTE 小区，圆圈的半径越大表示流量越高 (a) 高流量区域由 HTTACs 表示 (b) MNO1 的 HTTAC 仅显示 $n=100$ 个 LTE 小区 (c) 在高流量区域内划分 5GDA 边界

这类模式有助于网络规划工程师定位 5G 部署的高密度子区域 $A_m^g \subseteq A_m^{N_c}$ ，并如图 4(b,d,f) 所示为三个移动网络运营商进行区域划分。手动划分区域的值分别为 MNO1、MNO2 和 MNO3 的 $A_m^g = [23.69, 25.20, 23.39] \text{ km}^2$ 。划分出的区域 A_m^g 是网络规划的重要输入。

V. 结论

我们展示了如何利用从公共开放仓库提取的网络数据来辅助网络规划阶段。特别是，考虑了 OpenCell-ID 数据库，因为数百万个对应于运营商传统网络的样本可供研究目的公开使用。所选数据库的一个有趣方面是 LTE 标识符如何在处理原始数据后以多种方式协助获取相关信息。具体来说，为了形成数据库矩阵 D ，理解众包数据库及其特征，识别唯一小区和 TAC ID ($\mathcal{N}_m$ 和 $\mathcal{T}_m$) 并确定现有基础设施中每个运营商的最高流量 N_c 小区。

所提出的程序推断现有 eNB 小区中的网络基础设施和流量信息，这些信息用于确定 HTTAC 区域的 ID (H_m)。在确定的 HTTAC 区域内，交通模式指导 5GDA 划分前的界定。对于 5GDA，MNO 的财务限制通过 N_c 小区的价值来解决，这限制了要包含在目标区域内的小区数量。如结果所示， $A_m^{N_c}$ 地理区域的 N_c 小区揭示了目标区域内的交通模式。这些交通模式随后指导网络规划工程师对每个 MNO 最高流量密度的 5GDA (A_m^g) 进行适当和手动划分。正如解释的那样，通过数据了解现有的网络基础设施可以提供有关流量及对应区域内最高的订户密度的有用信息。网络数据采集实体的输出是划定的 5GDA (A_m^g)，作为网络尺寸设计实体的主要输入。

关于开放交通数据库，它们所包含的数据有许多潜在的应用。然而，在进行先前分析之前，必须仔细地对这些数据进行映射和使用，因为这些数据库是通过自愿贡献来填充的，并且不能保证其存储的原始数据的质量。本文描述了网络数据采集方法来解决这一问题。

致谢

REFERENCES

- [1] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," *2008 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 1–11, Oct 2008.
- [2] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, pp. 1147–1161, April 2017.
- [3] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Geospatial and temporal dynamics of application usage in cellular data networks," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 1369–1381, July 2015.
- [4] X. Cheng, L. Fang, X. Hong, and L. Yang, "Exploiting mobile big data: Sources, features, and applications," *IEEE Network*, vol. 31, no. 1, pp. 72–79, 2017.
- [5] S. Han, I. Lin, G. Li, S. Wang, and Q. Sun, "Big data enabled mobile network design for 5G and beyond," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 150–157, 2017.
- [6] X. Cheng, L. Fang, L. Yang, and S. Cui, "Mobile big data: The fuel for data-driven wireless," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1489–1516, 2017.
- [7] F. Malandrino, C. Chiasserini, and S. Kirkpatrick, "Cellular network traces towards 5G: Usage, analysis and generation," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 529–542, March 2018.
- [8] Z. Frias, L. Mendo, and E. J. Oughton, "Evolving mobile networks towards 5G. a framework to infer the state of connectivity infrastructure in dense urban areas," *30th ITS European Conference, Helsinki-Espoo, Finland*, pp. 1–24, June 2019.
- [9] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 901–911, 2016.
- [10] B. Yin and J. Lu, "A cost-efficient framework for crowdsourced data collection in vehicular networks," *IEEE Internet of Things Journal*, pp. 1–1, 2021.
- [11] J. Steenbruggen, E. Tranos, and P. Nijkamp, "Data from mobile phone operators: A tool for smarter cities?," *Telecommunications Policy*, vol. 39, no. 3, pp. 335–346, 2015.
- [12] K. Yadav, P. Kumaraguru, A. Goyal, A. A. Gupta, and V. Naik, "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, p. 1 – 6, 2011.
- [13] L. He, Z. Yan, and M. Atiquzzaman, "LTE/LTE-A network security data collection and analysis for security measurement: A survey," *IEEE Access*, vol. 6, pp. 4220–4242, 2018.
- [14] S. SchieSSL, T. Kopacz, and D. Heberling, "Towards a crowdsourcing-based EMF exposure monitoring: Evaluation of smartphone measurements using kriging," *2021 15th European Conference on Antennas and Propagation (EuCAP)*, pp. 1–5, 2021.
- [15] F. Malandrino, C. F. Chiasserini, and S. Kirkpatrick, "The impact of vehicular traffic demand on 5G caching architectures: A data-driven study," *Vehicular Communications*, vol. 8, pp. 13–20, 2017.
- [16] J. Cainey, B. Gill, S. Johnston, J. Robinson, and S. Westwood, "Modelling download throughput of LTE Networks," *39th Annual IEEE Conference on Local Computer Networks Workshops*, pp. 623–628, 2014.

- [17] Z. Frias, L. Mendo, and E. J. Oughton, “How does spectrum affect mobile network deployments? empirical analysis using crowd-sourced big data,” *IEEE Access*, vol. 8, pp. 190812–190821, 2020.
- [18] F. Malandrino, S. Kirkpatrick, and D. Bickson, “What is LTE actually used for? an answer through multi-operator, crowd-sourced measurement,” 2016.
- [19] “Exisiting 4G infrastructure from OpenCelliD by Uniwired Labs.” 2019. [Online]. Available: <https://opencellid.org/>.
- [20] “Spain’s 5G National Plan,” pp. 1–40, 2020.
- [21] Statista, “Mobile telephony market share in Spain by telecommunications providers.” 2019. [Online]. Available: <https://www.statista.com/statistics/745319/mobile-phone-provider-market-share-in-spain/>.
- [22] N. M. N. Operators, “Halber Bastion, Radio Frequency Technologies.” 2019. [Online]. Available: <https://halberdbastion.com/intelligence/countries-nations/spain>.