IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2020, accepted May 28, 2020, date of publication June 1, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998918

A Context-Aware Data-Driven Algorithm for Small Cell Site Selection in Cellular Networks

JUAN L. BEJARANO-LUQUE, MATÍAS TORIL, MARIANO FERNÁNDEZ-NAVARRO, ANTONIO J. GARCÍA, AND SALVADOR LUNA-RAMÍREZ, (Member, IEEE)
Department of Communication Engineering, University of Málaga, 29010 Málaga, Spain

Corresponding author: Juan L. Bejarano-Luque (jlbl@ic.uma.es)

This work was supported in part by the Horizon 2020 Project LOCUS through the European Union under Grant 871249, and in part by the Spanish Ministry of Economy and Competitiveness under Grant RTI2018-099148-B-I00.

ABSTRACT In mobile networks, detecting and eliminating areas with poor performance is key to optimize end-user experience. In spite of the vast set of measurements provided by current mobile networks, cellular operators have problems to pinpoint problematic locations because the origin of such measurements (i.e., user location) is not registered in most cases. At the same time, social networks generate a huge amount of data that can be used to infer population density. In this paper, a data-driven methodology is proposed to detect the best sites for new small cells to improve network performance based on attributes of connections, such as radio link throughput or data volume, in the radio interface. Unlike state-of-the-art approaches, based on data from only one source (e.g., radio signal level measurements or social media), the proposed method combines data from radio connection traces stored in the network management system and geolocated posts from social networks. This information is enriched with user context information inferred from traffic attributes. The method is tested with a large trace dataset from a live Long Term Evolution (LTE) network and a database of geotagged messages from two social networks (Twitter and Flickr).

INDEX TERMS Small cell, social network, twitter, traces, site selection.

I. INTRODUCTION

In the last years, mobile networks have experienced a continuous growth in the amount of users and services [1]. Likewise, the development of 5G system will increase drastically the number and heterogeneity of connected devices [2]. As a result, network complexity will make it very difficult for operators to manage their networks. For this reason, automation of mobile networks has become a field of interest for the industry and academia, giving rise to Self-Organizing Networks (SON) [3]. SON methods are classified into self-configuration, self-tuning and self-healing, depending on their use for network planning, optimization or problem solving.

At the same time, network providers have started to think in terms of users experience when managing their networks. Traditionally, network management followed a network-centric approach based on Quality of Service (QoS) criteria. This legacy approach has been replaced by a more user-centric approach based on how the user perceives the service, known as Quality of Experience (QoE) [4]. Parameterizing user experience for the different services provided by the network helps to increase the impact of network management on the end user. Thus, Customer Experience Management (i.e. CEM) is today one of the main procedures that stand out operators from their competitors [5]. Unfortunately, establishing the relationship between network performance and user opinion is a difficult task due to the large number of factors influencing user experience [6].

The communication context (e.g., terminal type, indoor/outdoor location, time of day, geolocation?) is one of the most influential factors on service perception [7], [8]. Thus, the most sophisticated QoE models take user context (e.g., time of day or user location) into account. To recognize user activity, service providers can use active measurements from on-body sensors through ensemble learning [9]. Alternative, network operators can infer user context by leveraging signaling events registered by the network on a per-connection basis [10]. In particular, indoor/outdoor detection can be performed based on signal level measurements [11] or

The associate editor coordinating the review of this manuscript and approving it for publication was Parul Garg.

traffic descriptors registered in connection traces [12]. Then, this information can be used to develop context-aware SON algorithms [13]–[15] [16].

To fulfill the stringent constraints of new use cases, 5G operators will take network performance to the next level by combining multiple techniques. Network densification has been recognized as an efficient way to provide higher network capacity and enhanced coverage [17]. In densely populated areas, densification is best achieved by combining macrocellular infrastructure with small cell (SC) deployments. In such heterogeneous networks, consisting of macrocells and small cells, discovering the best locations for SCs is key to making the most of the new infrastructure. However, most current site selection approaches only take simple network coverage and signal quality indicators due to the difficulty of modeling dynamic packet scheduling with users of multiple services and different radio link conditions in a radio network planning tool. In the absence of such a performance model, the Minimization of Drive Tests (MDT) feature [17], [18] allows the collection of geolocated measurements that can be used to build precise network performance maps (Radio Environment Map, REM). Such maps can then be used to detect coverage holes [19]. Unfortunately, MDT is rarely activated in live networks due to the workload of processing these measurements. Thus, network re-planning and optimization tasks often has to be done based on measurements only positioned by cell identity and time advance statistics. Such an approach leads to large location errors, which prevent estimating user context.

With recent advances in information technology, the interest in data science has grown in the last years. As a result, many open data initiatives have been launched around the world. Open data portals now offer direct and automated access to valuable assets that may be used to improve cellular network management. Some companies (e.g., OpenSignal [20] or WeFi [20]) provide real crowdsourced measurements collected by anonymous users, which can be used to assess current deployments [21]. Social networks are another source of information for understanding user behavior. Social media activity can be used to predict cellular traffic, regardless of radio access technologies or network providers [22]. At the same time, information on social events obtained from browser results or open data repositories can be used to explain abnormal network behavior during troubleshooting procedures [23]. Likewise, areas of poor signal coverage or service performance (i.e., blackspots) can be detected by processing geotagged text messages in social networks [24]. However, to the authors' knowledge, few works have considered the fusion of geolocated information from social networks and mobile networks.

In this work, a new data-driven site selection method for SCs is presented. The aim is to detect (and rank) small space regions (both indoors and outdoors) with lack of coverage or capacity with the largest benefit in terms of expected recovered data volume. Unlike legacy approaches, based solely on REMs derived with telecom data, the proposed method enriches this information with geotagged posts publicly available from social networks. Moreover, the method takes advantage of user context to refine user positioning. The main benefit is an improved spatial resolution in detecting inadequately served hotspots. Method assessment is performed with a real dataset consisting of a large set of open maps, connection traces from a live Long Term Evolution (LTE) network and geotagged posts obtained from Twitter [25] and Flickr [26]. The rest of this paper is structured as follows. Section II reviews the current state of research to clarify the contribution of this work. Section III describes the proposed SC site selection method. Section IV shows the results of the method in a live scenario. Finally, section V summarizes the main conclusions of the work.

II. SITE SELECTION RELATED WORK

The antenna placement problem (APP) can be formulated as a classical optimization problem. The design variables are the base station coordinates, often restricted to a limited set of candidate locations, and the objective function may be any combination of global network performance indicators. For computational reasons, this combinatorial optimization problem is often solved by heuristic approaches. Previous works can be classified by type of environment, design criteria and solution algorithm.

In terms of radio environment, a first group of works deal with macrocellular and microcellular outdoor scenarios. In [27], simulated annealing is used to solve the APP in a TDMA/FDMA microcell environment based on signal-to-interference-plus-noise ratio (SINR) and path loss indicators. In [28], APP in WCDMA is formulated as an integer linear programming problem, solved by tabu search. In [29], the APP is formulated so as to find the minimum number of antennas for a desired coverage level. In [30], the aim of the APP is to maximize coverage in GSM while still satisfying a minimum SINR requirement, which is achieved by genetic algorithms. In [31] and [32], a sensitivity analysis is carried out to check the impact of site location and antenna tilt angles on the pole capacity in a WCDMA network with uneven traffic distribution. In [33], randomized local search and tabu search are used to solve the APP in order to jointly optimize installation costs, signal quality and traffic coverage in WCDMA.

A second group of works extend the previous methods to indoor scenarios. In some of them, the APP is formulated to minimize path loss (or maximize coverage) with a general-purpose optimization algorithm (e.g., genetic [34], direct search [35], simulated annealing [36] or heuristic [37]). Similarly to [29], [38] proposes binary integer programming to find the minimum number of access points guaranteeing a minimum SINR in the scenario. In [39], a heuristic method is proposed to place indoor access points in WCDMA with constraints on uplink (UL) and downlink (DL) SINR. Later studies focus on SINR optimization by different methods (e.g., brute force enumeration in WCDMA [40] and LTE [41], particle swarm in WCDMA [42] and reduction approximation in

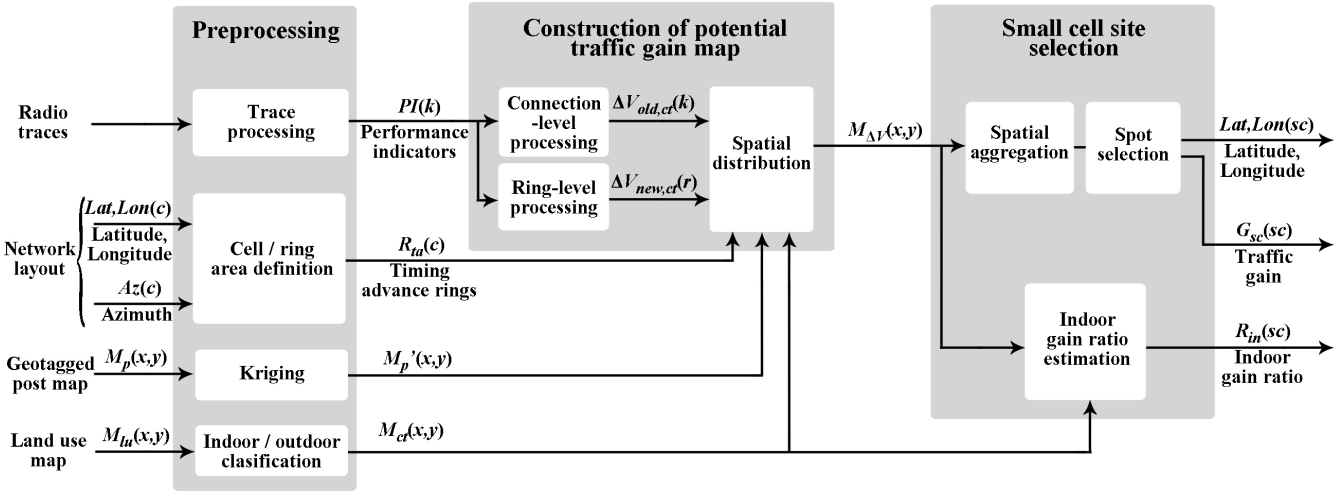


FIGURE 1. **Block diagram of the proposed Trace and Social Method (TSM) for small-cell site selection.**

WCDMA [43]). In [44], a method for femtocell placement is proposed to minimize transmit power of mobile users. In [45], a methodology for locating enterprise femtocells in a building is proposed to maximize the effectiveness of mobility load balancing schemes. Alternatively, in [46], [47], the aim is to find the best location for Wi-Fi access point for optimal user positioning in indoor environments. More recently, an overview of SC deployment strategies for Internet of Things (IoT) 5G environments is presented in [48].

Social networks can provide valuable information for the APP. In [22], it is shown that social media activity aggregated at a district level can be used to predict cellular traffic at a spatiotemporal resolution higher than current approaches based on census data. Equally important, social media data reflects the overall traffic demand across radio access technologies or network providers. Similarly, cell loads can be estimated by a queuing model adjusted with the distribution of geotagged messages from Twitter [49]. Nonetheless, it is still to be checked whether data from social networks can be used to predict traffic at a lower scale (e.g., at a building level).

The main contributions of the method proposed here are: a) to combine cellular network measurements with user context and social media data to detect blackspots at a building level, and b) to quantify the expected performance benefit of each new SC based on observed user behavior.

III. METHODOLOGY

The aim of the method is to detect areas in a live cellular network where traffic demand can be increased by improving radio link conditions. For this purpose, a deep knowledge of the most influential factors affecting traffic demand is needed, namely: a) radio network performance per location, b) spatial user distribution and c) context-dependent user behavior.

Fig. 1 shows a block diagram of the proposed method. The inputs are: a) radio trace files, comprising signaling events of individual connections in the radio interface, b) the cellular network layout, describing existing site coordinates and antenna azimuths in the area, c) a large dataset of geotagged messages (posts) generated in the area from social networks, and d) a land use map of the area. This inputs are used due to its relevant information. In this way, radio traces provides all the necessary information about network performance, while the land uses map adds context information and the network layout and the dataset of geotagged posts allows to build a deep knowledge of the spatial user distribution. On the one hand, the network performance enables to compute the expected traffic gain. On the other hand, spatial user distribution and context information determine where this gain occurs. The output is an ordered list of candidate SC sites, specified by their geographical coordinates (latitude and longitude), $Lat/Lon(sc)$, expected traffic volume gain, $G_{sc}(sc)$, and ratio of traffic gain due to indoor locations, $G_{in}(sc)$.

In a first pre-processing stage, connection traces are analyzed to compute performance indicators for each connection, $PI(k)$ (k denotes connection index). Likewise, connections are also analyzed in groups based on location. To this end, connections are positioned based on the combination of cell identity and timing advance information (known as Enhanced Cell ID, ECID). Timing Advance (TA) is a temporal offset introduced at the terminal to ensure that the downlink and uplink subframes are synchronized at the base station. Such an offset takes discrete values depending on the distance between user and base station. TA statistics are collected on a cell basis, which are used to divide cell service areas into concentric distance rings centered at the base station. With this information, connections are grouped per ring, $R_{ta}(c)$ (R for ring, c denotes cell index and ta denotes distance ring), to perform ring-level analysis. The spatial distribution generated from discrete locations of individual geotagged posts is

interpolated by a kriging algorithm to generate a continuous map approximating the density of users of social networks, $M'_p(x, y)$ (x and y are Cartesian coordinates). Finally, the land use map is simplified into a map showing indoor and outdoor locations, $M_{ct}(x, y)$ (ct for context).

In a second stage, a map is derived showing the potential traffic gain achieved at each point if a new SC was deployed, $M_{\Delta V}(x, y)$. In the live network, such an increase is mainly originated by two effects: a) the increase of session length (e.g., more downloaded web pages, larger time of audio playback?) or content quality (e.g., higher resolution video) due to a better user experience, and b) the request of more data-hungry services (e.g., videostreaming, app download?). In this work, the former component, related to session length, is associated to past connections already established without the SC (referred to as old connections), whereas the latter component, related to new services, is associated to fresh connections that would be established with the new SC (referred to as new connections). Moreover, gain values are different depending on user context (indoors or outdoors). From these assumptions, the potential traffic gain for existing connections per context, $\Delta V_{old,ct}(k)$, is computed by processing past connections individually. In contrast, the potential traffic gain for new connections, for which no data is available, is computed at a ring level, $\Delta V_{new,ct}(r)$, as explained later. Then, traffic gains are distributed in space using the spatial user distribution inferred from post messages.

In a third stage, traffic gains per tile are aggregated at a cell level by taking into account typical small cell radii. Finally, the best SC locations are selected based on the expected traffic gain. To ease the understanding of the algorithm, Table 1 contains the most important notation in the text. Unless stated otherwise, all variables refer to the Downlink (DL).

TABLE 1. Notation table.

Notation	Description	Units
$V(k)$	Data volume generated by connection k	[bytes]
$T(k)$	Throughput of the connection k	[kbps]
$R_{nlt,ct}(r)$	Ratio of connections non-last TTI vs last-TTI connections in context ct and distance ring r	[-]
$\Delta V_{opt,ct}(k)$	Data volume gain from existing non-last TTI connection k, generated in context ct, due to better radio link performance	[bytes]
$\Delta V_{new,ct}(r)$	Data volume gain from new non-last TTI connections in context ct of the ring r, due to better radio link performance	[bytes]
$\Delta V_{opt}(x, y)$	Map of data volume gain obtained from existing non-last TTI connections	$\left[\frac{\text{bytes}}{\text{tile}}\right]$
$\Delta V_{new}(x, y)$	Map of data volume gain obtained from new non-last TTI connections	$\left[\frac{\text{bytes}}{\text{tile}}\right]$
$M_{\Delta V}(x, y)$	Map of potential data volume gain in the network	$\left[\frac{\text{bytes}}{\text{tile}}\right]$
$G_{sc}(x, y)$	Map of data volume gain obtained from a small cell in tile (x,y)	$\left[\frac{\text{bytes}}{\text{tile}}\right]$
$R_{in}(x, y)$	No. of power limited measurements in UL	[-]

TABLE 2. Selected performance indicators.

DL throughput	[kbps]
DL volume	[bytes]
UL volume	[bytes]
RRC connection duration	[s]
No. of active time transmission intervals (TTIs)	[ms]
RSRP histogram	[dBm]
DL CQI histogram	[-]
UL SINR in PUCCH	[dB]
DL spectral efficiency	$\left[\frac{\text{b}}{\text{RE}}\right]$
UL spectral efficiency	$\left[\frac{\text{b}}{\text{RE}}\right]$
No. of measurements in Rank 1	[%]
No. of measurements in Rank 2	[%]
No. of no power limited measurements in UL	[%]
No. of power limited measurements in UL	[%]

A. INPUT DATA

Radio traces are log files with signaling events generated by base stations, which are periodically uploaded to the mobile network management system. This data is delivered as binary files that must be decoded to extract performance measurements at a connection level. The reader is referred to [50] for details on trace processing. Table 2 summarizes the data fields needed by the algorithm.

The data fields used in this method to obtain the potential traffic gain in each point of the network are DL throughput, DL volume and the probability of each connection happening indoor. This probability is obtained with the rest of indicators presented in Table 2, plus DL throughput, as described in [12].

The geotagged posts are obtained from two social networks (Twitter [25] and Flickr [26] by Application Programming Interfaces (APIs) provided by the service provider [51]. Each post is collected with an associated location, which is used to build a raster (i.e., grid-based data) with the number of posts per location.

The map of land uses consists of a raster representing the business and social activity of each small piece of terrain (tile). This information is publicly accessible from open data initiatives fostered by institutions (e.g., local municipality) or popular crowdsourcing platforms (e.g., OpenStreetMap [52]).

B. CONSTRUCTION OF POTENTIAL TRAFFIC GAIN MAP

First, the basis of the method is introduced. Then, a preliminary analysis over real traces proves the validity of the proposed approach by showing how radio link conditions affect traffic generated by mobile users. Finally, the algorithm to compute the potential traffic gain per location is detailed.

1) RATIONALE

The deployment of a new SC can increase signal level (coverage), signal quality (spectral efficiency) and, ultimately, link capacity perceived by the user. It is expected that the performance gain obtained by these changes will be much larger for users of data-intensive services (e.g., videostreaming, app

download?) than for users of low data volume services (e.g., instant messaging, voice call?). Unfortunately, connections in current radio access networks are roughly divided into large groups of services. In the absence of a precise classification of services, last-TTI transmission statistics are used here to identify data-intensive services. A last TTI is the last transmission interval of a data burst that temporarily empties the transmit buffer [53]. Thus, data-intensive services, consisting of one or more large data bursts, often have low last-TTI ratios. In contrast, non-data-intensive services, consisting of one or more small data bursts, have large last-TTI ratios. For these reasons, data-intensive services tend to behave as a full-buffer traffic source, whereas non-data intensive services can be modeled as bursty traffic.

Based on the above observations, two main sources of traffic gain are identified. A first component of traffic gain comes from users already using data-intensive (i.e., non-last TTI) services without the SC, which will increase their traffic as a result of the increased session length and content quality due to a better user experience. A second source of traffic gain comes from users that, in the past only requested non-data intensive (i.e., last-TTI) services, but with the SC request new data-intensive services due to the better user experience. In addition, since user behavior is not the same in all locations, both traffic components are broken down depending on user context (indoors or outdoors). The convenience of this approach is validated with the analysis presented next.

2) IMPACT OF RADIO LINK THROUGHPUT ON USER TRAFFIC

A priori, it is difficult to predict the impact of adding a new SC on user traffic. To this end, a correlation analysis is carried out based on traces collected in a live LTE network. The aim is to model the relationship between radio link throughput and several indicators related to traffic demand.

To check how the data volume of users of non-last TTI services is affected by network conditions, the relationship between data volume and user throughput is studied on a per-connection basis. Fig. 2 shows a scatter plot of these two variables observed in real mobile connections (1 point per connection). The throughput value on the x-axis corresponds to the average throughput of the connection at Packet Data Convergence Protocol (PDCP) layer excluding last TTIs (same for data volume on the y-axis) [54]. Since the focus is on non-last TTI services, only connections with a last-TTI volume ratio lower than 10% are considered in the analysis (i.e., full-buffer connections). For clarity, three regression curves are included in the figure. The middle one is obtained by linear regression on all the samples (note the log scale in the x-axis), and hence reflects the general trend. The other two are computed by quantile regression to reflect the trend of extreme cases. Specifically, the lower/upper curve is obtained by linear regression on the 10th/90th percentile of data volume of connections in different throughput ranges. In the three curves, the data volume per connection increases with average user throughput, which proves the hypothesis that improving radio link conditions generally leads to higher

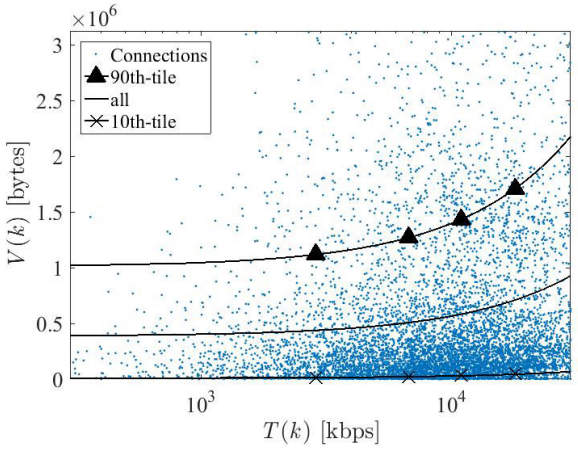


FIGURE 2. Example of scatter plot of data volume vs radio link throughput per connection.

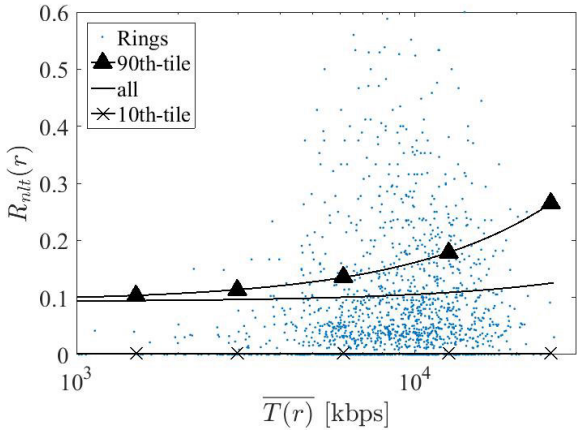


FIGURE 3. Example of scatter plot of ratio of non-last TTI connections vs radio link throughput per timing advance ring.

data volumes per connection. As expected, correlation is not strong, since two connections may have very different data volumes for the same throughput value due to different session lengths. This observation justifies the need for several regression curves to model the relationship between throughput and data volume per connection. A closer analysis of residuals shows that the mean absolute error is reduced from 4.93e6 with a single curve to 2.25e6 with 3 curves (achieving a reduction of 45.64 %).

To check if the number of requests of non-last TTI services of a user is affected by network conditions, the correlation between the ratio of non-last TTI against last-TTI connections is examined. As above, a connection is considered as a non-last TTI (last TTI) connection if less than 10% of data is transmitted in last TTIs (non-last TTIs). All other connections that fall in between these two thresholds are discarded in the analysis. As explained above, in the absence of geolocated traces, connections are positioned based on cell identity and timing advance. Thus, the analysis can be done by aggregating all connections within a TA ring. For reliability, only rings with more than 25 connections are considered. Fig. 3 shows a scatter plot of the non-last TTI/last TTI connection ratio

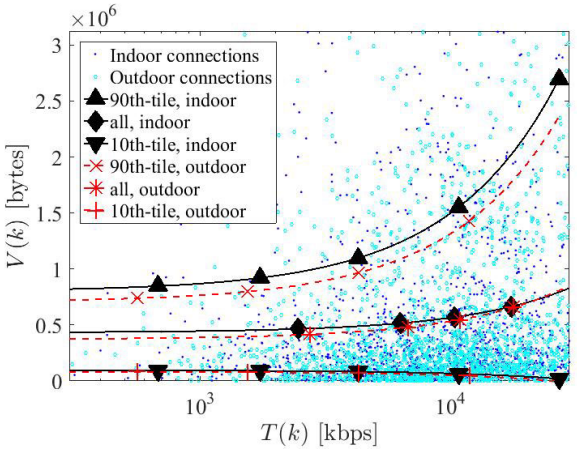


FIGURE 4. Example of scatter plot of data volume vs radio link throughput per connection, broken down by context.

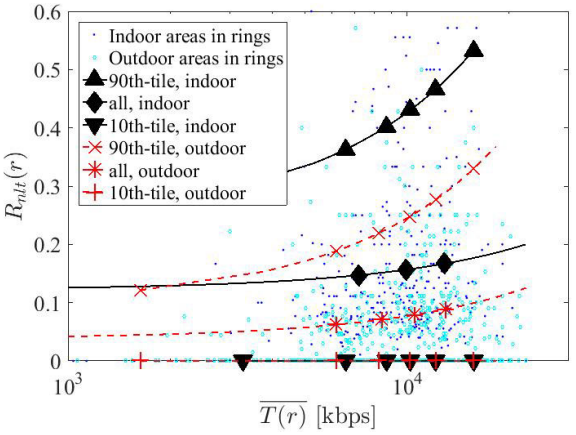


FIGURE 5. Example of scatter plot of ratio of non-last TTI connections vs radio link throughput per timing advance ring, broken down by context.

versus average connection throughput in the ring (1 point per cell/ring/context). Again, three regression curves are superimposed, obtained by linear and quantile regression. In the general and 90th-tile trend curves, it is confirmed that rings with a larger average throughput tend to have larger ratios of non-last TTI connections. Likewise, the large dispersion justifies the need for several regression curves also in this case (in this case, the mean absolute error with 3 curves is reduced by 53.28 %).

The previous analyses can be done separately for indoor and outdoor connections to check the impact of user context. In the absence of a precise user positioning method, connections are tagged as indoor or outdoor based on trace measurements as described in [12]. Only connections with a large confidence of being indoors or outdoors (>90%) are considered in the analysis. Fig. 4 and 5 present the results of breaking down the regression curves per context for the two considered traffic indicators. In Fig. 4, the outdoor and indoor curves differ slightly for connection data volume, which points out that data volume per connection is not much affected by user context. However, in Fig. 5, indoor curves are well over outdoor curves, indicating that, for the same user throughput conditions, the ratio of data-intensive vs non-data-intensive services is larger indoors than outdoors. This observation justifies the need for a different set of regression curves for indoors and outdoors.

From the previous analysis, two regression models are derived, each consisting of 6 regression curves. A first model defines the relationship between data volume, V, and average connection throughput, T, in a non-last TTI connection k that took place in context ct as

$$V^{(j)}(k) = \beta^{(j)}_{1,V,ct} T(k) + \beta^{(j)}_{0,V,ct} \quad k \in ct, \tag{1}$$

where $\beta_{i,V,ct}(j)$ is the ith regression coefficient of the regression curve j ($j \in \{10th, all, 90th\}$) and user context ct ($ct \in \{in, out\}$ for indoors and outdoors, respectively). Regression $j = all$ is a simple linear regression with all the samples, while j=10th and 90th correspond to quantile regression with 10th and 90th percentile of data volume values per throughput bin, respectively. A second model defines the relationship between the ratio of non-last TTI vs last-TTI connections, R_{nlt}, and the mean throughput from all connections, $\overline{T}$, in a TA ring r in context ct as

$$R^{(j)}_{nlt,ct}(r) = \beta^{(j)}_{1,R,ct} \overline{T}_{ct}(r) + \beta^{(j)}_{0,R,ct}, \tag{2}$$

where $\beta^{(j)}_{i,R,ct}$ is the i^{th} regression coefficient for the regression curve j and user context ct.

3) ESTIMATION OF POTENTIAL TRAFFIC GAIN PER CONNECTION AND RING

The traffic gain obtained by deploying a new SC is divided in two components: a) the increase from existing connections, calculated on a connection basis (only applicable to connections that needed many resources, i.e., non-last TTI connections), and b) the increase from the extended use of data-intensive services, calculated on a ring basis.

As a first step, an optimal throughput value is defined, T_{opt}, as the best throughput that can be reached with optimal propagation conditions. In this work, T_{opt} is set to the 90th percentile of average throughput across all connections in the network.

Then, connections are classified as indoor or outdoor. To this end, the probability that a particular connection is generated in a given context, $P_{ct}(k)$, is estimated from trace data with the algorithm described in [12]. As a result, a connection might be tagged both as indoor and outdoor, provided that $P_{in}(k) + P_{out}(k) = 1$.

Following the above classification, the potential gain of data volume from an existing connection is divided into an outdoor and indoor component, calculated as

$$\Delta V_{opt,ct}(k) = P_{ct}(k)(V'_{ct}(k) - V(k)), \tag{3}$$

where $P_{ct}(k)$ is the indoor/outdoor probability of the connection, $V(k)$ is the data volume of the connection, taken from traces, and $V'_{ct}(k)$ is the potential data volume indoors/outdoors of the connection with optimal radio conditions (i.e., after deploying the SC). The latter is estimated

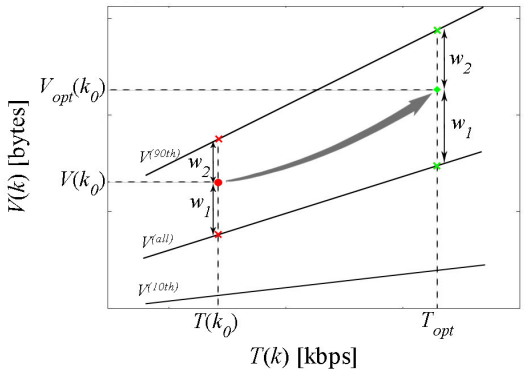


FIGURE 6. **Estimation of $V_{opt}(k)$.**

by the regression model in (1) with $T(k) = T_{opt}$. Yet, the specific regression curve (out of the three possible, j=10th, all or 90th) must be decided. To solve this problem, linear interpolation between the closest curves is used, as shown in Fig. 6. Specifically,

$$V(k) \geq \widehat{V_{90}}(T(k)) \left\{ w_1 = 1 \right. \tag{4}$$

$$\widehat{V_{all}}(k) \geq V(k) > \widehat{V_{90}}(T(k)) \left\{ \begin{array}{l} w_1 = \frac{V(k) - \widehat{V_{all}}(T(k))}{\widehat{V_{90}}(T(k)) - \widehat{V_{all}}(T(k))} \\ w_2 = 1 - w_1 \end{array} \right. \tag{5}$$

$$\widehat{V_{10}}(t) \geq V(k) > \widehat{I_{all}}(T(k)) \left\{ \begin{array}{l} w_1 = \frac{V_{ct} - \widehat{V_{10}}(T(k))}{\widehat{V_{all}}(T(k)) - \widehat{I_{10}}(T(k))} \\ w_2 = 1 - w_1 \end{array} \right. \tag{6}$$

$$V(k) < \widehat{V_{10}}(T(k)) \left\{ w_1 = \frac{V(k)}{\widehat{V_{10}}(T(k))} \right. \tag{7}$$

To reduce the influence of outliers, input data volume is upper and lower bounded by the values in the 90th and 10th percentile curves.

The potential increase in data volume from the use of more data-intensive services is estimated per ring from the number of connections of non-data intensive services, which should not be affected by radio link conditions, and is thus the same with and without the SC. Specifically, the increase in the number of ideal non-last TTI connections (i.e., those transmitting all its data in non-last TTIs) per ring and context is estimated as

$$\begin{aligned} \Delta N_{new,ct}(r) &= N'_{nlt,ct}(r) - N_{nlt,ct}(r) \\ &= N_{lt,ct}(r) R'_{nlt,ct}(r) - N_{nlt,ct}(r), \end{aligned} \tag{8}$$

where $N_{nlt,ct}(r)$ and $N_{lt,ct}(r)$ are the number of equivalent non-last TTI and last-TTI connections per context ct in ring r without the SC, calculated as

$$N_{nlt,ct}(r) = \sum_{k \in r} P_{ct}(k) R_{nlt}(k), \tag{9}$$

$$N_{lt,ct}(r) = \sum_{k \in r} P_{ct}(k)(1 - R_{nlt}(k)), \tag{10}$$

where $R_{nlt}(k)$ is the share of data volume transmitted in non-last TTIs in connection k (e.g., 4 connections with $P_{ct}(k) = 0.5$ and $R_{nlt}(k) = 0.5$ are equivalent to 1 ideal non-last TTI connection). Likewise, $N'_{nlt,ct}(r)$ is the predicted number of ideal non-last TTI connections with the SC, computed from the new non-last-TTI/last-TTI connection ratio with the SC. The latter is estimated by the regression model in (2) with the curve interpolation process explained in Fig. 6 and $\overline{T}_{ct}(r) = T_{opt}$. Then, the increase in data volume in ring r is calculated as

$$\Delta V_{new,ct}(r) = \Delta N_{new,ct}(r) \overline{V_{opt,ct}}(r), \tag{11}$$

where $\overline{V_{opt,ct}}(r)$ is the average data volume of an ideal non-last TTI connection in ring r with context ct with the new SC, calculated as

$$\overline{V_{opt,ct}}(r) = \frac{\sum_{k \in r,ct} V'(k)}{N_{nlt,ct}(r)}. \tag{12}$$

C. SPATIAL DISTRIBUTION

Once the two traffic gain components are calculated on a per-connection and ring basis, these have to be projected onto the map by positioning connections. Unfortunately, connections in traces are rarely geolocated, so they must be located by ECID method. This leads to large positioning errors in rings far from the serving cell. To circumvent this problem, the spatial user distribution within a ring can be inferred from the distribution of geotagged posts taken from social networks as in [22], since the transmission of short messages is not conditioned on a good radio link.

The geolocation process starts by creating a grid with the same tile dimensions as the map of land uses, $M_{lu}(x, y)$. Hereafter, indexes (x, y) refer to horizontal and vertical tile indexes. From land uses, a user context matrix with the same size, $M_{ct}(x, y)$, is derived indicating whether every tile is indoor or outdoor. Then, a matrix with the number of geolocated posts per tile, $M_p(x, y)$, is constructed by taking advantage of location information provided by social networks. In areas of low population density (e.g., open field), the average number of posts per tile is much lower than one, causing that most tiles have no posts and a few of them have some. For a better estimation of these small density values, a kriging process [55] is applied to drive the underlying spatial post distribution, $M'_p(x, y)$.

The post distribution is used to derive the probability of a connection occurring in a tile (x, y) labeled as context ct (i.e., indoor or outdoor) in ring r as

$$P(r, x, y) = \frac{1 + M'_p(x, y)}{N_{ct}(r) + \sum_{(x,y) \in r} M'_p(x, y)} \quad \begin{array}{l} (x, y) \in r, \\ (x, y) \in ct, \end{array} \tag{13}$$

where $N_{ct}(r)$ is the number of tiles labeled as context ct in ring r and $M'_p(x, y)$ is the post spatial distribution. Note that the same tile index (x, y) can be served by rings from different cells and, thus, different $P(r, x, y)$ values are associated to each ring serving the same tile. A closer analysis of (13) shows that, in rings where the number of geotagged posts is 0 (as could be in unpopulated areas), connections registered in traces are uniformly distributed in the ring area (i.e., $P(r, x, y) = 1/N_{ct}(r)$). In contrast, in rings with a

large number of geotagged posts, connections are distributed following the post distribution in the ring (i.e., $P(r,x,y) \approx (M'_p(x,y))/sum_{(x,y)\in r} M'_p(x,y)$).

Once tile probabilities are calculated, traffic gains are projected onto the map. To this end, the data volume increase from existing connections (per connection) or new connections (per ring) is distributed in indoor/outdoor tiles within the ring according to probabilities in (13) and then aggregated across rings serving the same tile. Specifically, the data volume increase due to existing or new connections in tile (x,y), $\Delta V_{opt}(x,y)$ and $\Delta V_{new}(x,y)$, respectively, is computed as

$$\Delta V_{opt}(x,y) = \sum_{r/(x,y)\in r} \sum_{k\in r} \Delta V_{opt,ct}(k) P(r,x,y) \quad (x,y) \in ct, \tag{14}$$

$$\Delta V_{new}(x,y) = \sum_{r/(x,y)\in r} \sum_{k\in r} \Delta V_{new,ct}(r) P(r,x,y) \quad (x,y) \in ct. \tag{15}$$

Finally, both terms are added to build the traffic gain map as

$$M_{\Delta V}(x,y) = \Delta V_{opt}(x,y) + \Delta V_{new}(x,y)\,. \tag{16}$$

D. SMALL CELL SELECTION

Once potential traffic gain map is obtained, it must be decided where the SC should be located. To this end, the aggregated data volume gain associated to a candidate SC location (x,y), $G_{sc}(x,y)$, is calculated as

$$G_{sc}(x,y) = \sum_{(i,j)\in A_{sc}(x,y)} M_{\Delta V}(i,j), \tag{17}$$

where $A_{sc}(x,y)$ is the coverage area of a hypothetical SC located in (x,y). In this work, SC is assumed to be omnidirectional, so that $A_{sc}(x,y)$ is a circle of radius 50 meters centered at (x,y) [56], [57].

Then, the best candidate SC locations can be identified by detecting local maxima in $G_{sc}(x,y)$ in an iterative manner. To avoid selecting too close sites that might overlap, the traffic gain map is updated every time a new SC is selected by forcing to 0 the value of points under the newly selected SC. The selection process is summarized as:

The output of the above process is a list of new SC sites specified by the tile including the center of their targeted service areas, with the.

Note that, in heterogeneous areas comprising indoor and outdoor tiles next to each other, the context of the center tile may not be the same as the context where most of the traffic gain comes from. For instance, the suggested SC location for a SC covering four nearby buildings (where most of the traffic gain comes from indoor users) might be on a street level (outdoors). To avoid this situation, a ratio of indoor improvement for each SC, $R_{in}(x_{sc},y_{sc})$, is estimated as

$$R_{in}(x_{sc},y_{sc}) = \frac{\sum_{(x,y)\in A_{sc}(x_{sc},y_{sc}),indoor} M_{\Delta V}(x,y)}{\sum_{(x,y)\in A_{sc}(x_{sc},y_{sc})} M_{\Delta V}(i,j)}. \tag{18}$$

Lower values indicate that most of the traffic gain comes from outdoor areas, and higher values indicate that the traffic gain comes from indoor areas. In the former case, the SC should be located outdoors, while, in the latter, the SC should be located indoors. In this work, a threshold of $R_{in,th}) = 0.5$ is heuristically set to decide if a SC must be located indoors (i.e., $R_{in}(x_{sc},y_{sc}) \geq 0.5$) or outdoors (otherwise). If the tile of the suggested SC location does not match that context, a closer analysis is needed to determine the best location according to the prevailing gain context. Such an analysis might end up with the addition of several SCs to cover indoor and outdoor locations separately.

Algorithm 1 Algorithm for Site Selection

Input: Map of aggregated data volume gain associated to a candidate SC location ($G_{sc}(x,y)$), minimum acceptable gain (G_{min}) and map of potential data volume gain per tile ($M_{\Delta V}(x,y)$)

Output: List of new SC (x_{sc},y_{sc}) and achieved gain ($G_{sc}(sc)$)

while $\max(G_{sc}(x,y)) > G_{min}$ **do**

\\Obtain the position of maximum gain:

$(x_{sc},y_{sc}) = \arg\max_{(x,y)} (G_{sc}(x,y))$

\\Obtain maximum gain:

$G_{sc}(sc) = \max(G_{sc}(x,y))$

\\Set to 0 all tiles covered by the new SC:

$M_{\Delta V}(x,y) = 0 \quad \forall (x,y) \in A_{sc}(x_{sc},y_{sc})$

\\Recompute map of SC volume gain:

$G_{sc}(x,y) = \sum_{(i,j)\in A_{sc}(x,y)} M_{\Delta V}(i,j) \quad \forall (x,y)$

end while

IV. METHOD ASSESSMENT

The proposed method is tested with a large set of traces obtained from a live LTE network. The assessment methodology is described first. Results are presented later, including an explanation of how the algorithm works with real data and a comparison with legacy approaches. Finally, computational aspects are discussed.

A. ANALYSIS SET-UP

The considered scenario covers a geographical area of 125 km^2, corresponding to the metropolitan area of a city with 800,000 inhabitants. This area is divided into tiles of 10×10 m. The land use per tile, $M_{lu}(x,y)$, is obtained from open data provided by the municipality. Table 3 shows the distribution of land uses in the scenario, with a brief description, their context classification (indoor or outdoor) and their share in the scenario.

The analyzed area comprises 400 LTE cells, grouped into 175 tri-sectorized sites, with a carrier frequency of 2.325 MHz and a system bandwidth of 15 MHz. In these cells, trace collection is activated for 2 hours, obtaining 166,561 connections. Table 4 shows the main statistics for the indicators derived from traces, namely radio link throughput, connection data volume and indoor probability.

The geotagged posts from social networks generated in the area are collected in real time during 16 months for

TABLE 3. Description of land uses in the scenario.

Land Use	Description	Indoor/Outdoor	Share
Services	Cultural, institutional, educational or medical services.	Indoor	10%
Offices	Office buildings for management, information or professional services.	Indoor	3%
Mixed	No clear use, mixture of many land uses.	Indoor	11%
Residential	Houses, hotels and visitor services.	Indoor	15%
Retail	Retail or entertainment.	Indoor	4%
Industrial	Industrial and manufacturing services.	Indoor	3%
Open Space	Fields and green areas.	Outdoor	22%
Paths	Ways, fields under Right of Way and paths, both pedestrians and used by vehicles.	Outdoor	30%
Rivers/Lakes	Inland water.	Outdoor	2%

TABLE 4. Statistics of input key performance indicators.

Source	10th percentile	Mean	90th percentile
T(k) [kbps]	0	9073	28648
V(k) [bytes]	173	5.68e5	6.28e5
$P_{in}(k)$ [%]	0.027	0.525	0.978

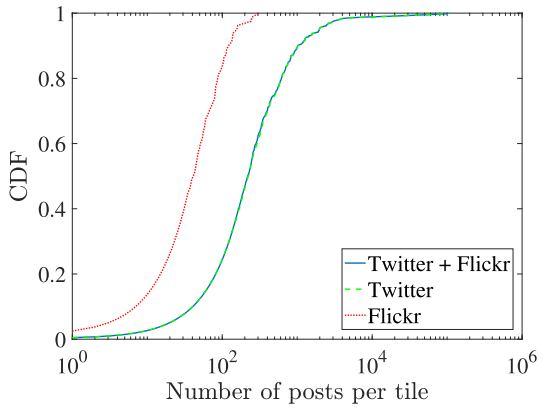


FIGURE 7. Spatial distribution of posts per social network.

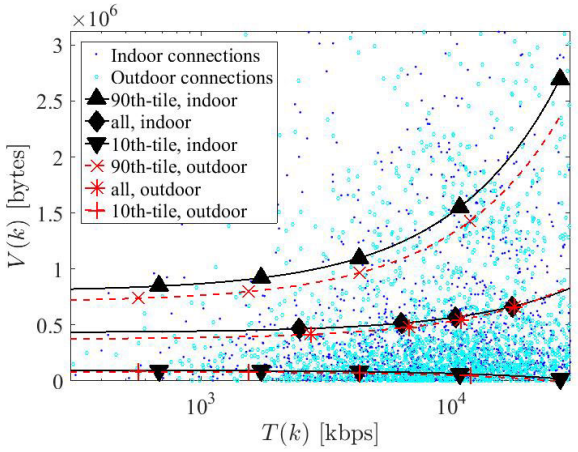


FIGURE 8. Scatter plot of data volume vs radio link throughput per connection, broken down by context.

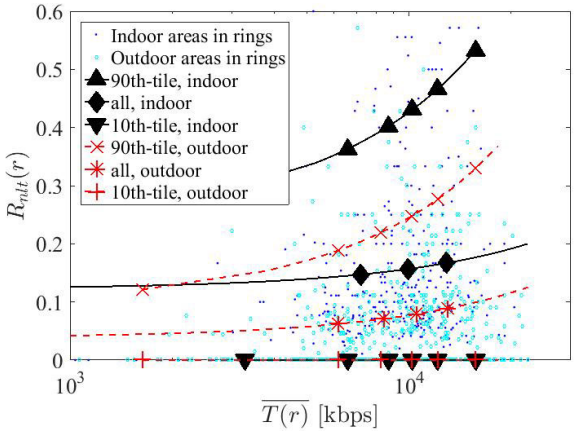


FIGURE 9. Scatter plot of ratio of non-last TTI connections vs radio link throughput per timing advance ring, broken down by context.

Twitter and 12 months for Flickr, resulting in 785,515 and 33,519 posts, respectively. The code used for this purpose is publicly available at [58]. Fig. 7 shows the Cumulative Distribution Functions (CDFs) of the number of posts in the scenario, $M_p(x, y)$, broken down per application (dashed line for Twitter, and dotted line for Flickr). It is observed that most posts come from tweets in Twitter.

Three methods are tested: a) the proposed method to select the best SC candidate locations, which combines traces and social network data (referred to as trace and social network method, TSM), b) a simplified version of the method based only on traces, which segregates connections into outdoor/indoor connections to refine the spatial user distribution from timing advance with the land use map (referred to as trace method, TM), and c) a variant of the method, inspired in the method to detect traffic hotspot proposed in [22], that derives the user spatial distribution only from social network data (referred to as social network method, SM). The latter two are legacy approaches used as benchmarks to check the benefit of only using posts or traces, respectively.

Ideally, method performance should be evaluated by deploying the sites suggested by the methods in the real network. Since this was not possible, the comparison presented here is based on the potential traffic gain map per tile derived by TSM. Thus, the analysis only shows where (and to what extent) the three methods behave differently.

B. METHOD PERFORMANCE

The aim here is to describe the results of the different stages in the TSM method, shown in Fig. 1.

ESTIMATION OF POTENTIAL TRAFFIC GAIN MAP

Fig. 8 and 9 show the scatter plot of data volume and non-last TTI connection ratio versus throughput built from the trace

TABLE 5. Regression parameters.

Context, ct		indoor			outdoor	
Percentile, j	90th	mean	10th	90th	mean	10th
$\beta^{(j)}_{0,V,ct}$	9.41e9	3.46e6	5.32e4	7.35e6	2.97e6	7.59e4
$\beta^{(j)}_{1,V,ct}$	284.21	105.81	21.98	365.96	127.89	12.76
$\beta^{(j)}_{0,R,ct}$	0.241	0.122	0	0.097	0.038	0
$\beta^{(j)}_{1,R,ct}$	1.86e-05	3.47e-06	0	1.48e-05	3.89e-06	0

TABLE 6. Statistics of data volume gain per existing connection and ring (in bytes).

Source	10th percentile	Mean	90th percentile
$\Delta V_{opt,in}(k)$	0	6.077e4	5.450e7
$\Delta V_{opt,out}(k)$	0	1.279e5	1.304e8
$\Delta V_{new,in}(r)$	0	7.104e6	1.980e9
$\Delta V_{new,out}(r)$	0	4.860e6	6.430e8

TABLE 7. Estimation of potential data volume increase (in bytes).

KPI	10th percentile	Median	90th percentile
$M_{\Delta V}(x,y)$	0	1.030e4	1.903e8
$M_{\Delta V_{opt}}(x,y)$	0	2.309e3	6.839e7
$M_{\Delta V_{new}}(x,y)$	0	3.187e3	1.818e8

dataset, which are used to derive the set of regression curves per context for each indicator. Table 5 presents the resulting regression coefficients.

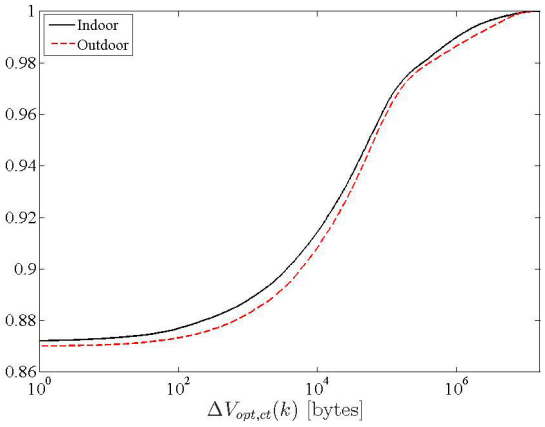
Once regression models are derived, the potential data volume gain per connection and ring are estimated. Table 6 summarizes the results presenting the 10th percentile, mean and 90th percentile values of data volume gains of connections and rings in the scenario, $\Delta V_{opt,ct}(k)$ and $\Delta V_{new,ct}(r)$. Fig. 10 shows their CDFs, broken down by context. As expected, in Table 6, it is observed that gain values are smaller for individual connections than for rings aggregating several locations. More interestingly, from Fig. 10(a), it is deduced that, for existing connections, larger data volume gains per connection are achieved outdoors. In contrast, from Fig. 10(b), it is deduced that indoor rings show a larger increase of new data-intensive connections.

C. PERFORMANCE COMPARISON

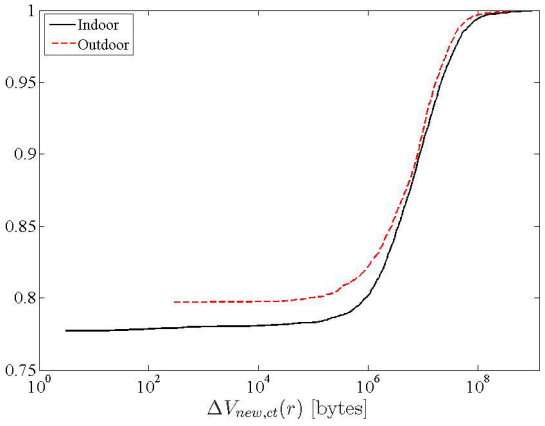
1) SPATIAL DISTRIBUTION

The potential gains estimated on a connection and ring basis are projected onto a map by geolocating network data. Fig. 11 shows the statistical distribution of several indicators computed per tile by aggregating connections in the tile, namely number of connections, average radio link throughput, total data volume and mean data volume per connection. The large variability observed in all indicators is just the consequence of the heterogeneity of the scenario, comprising areas of very different population density and radio link conditions (e.g., the number of connections per tile of 100 m^2 ranges from 0 to 175 connections). This justifies the need for a precise model that considers all the above factors.

Fig. 12 (a)-(c) depict the potential traffic gain map, broken down by its two components, over the orthophoto of the area. From left to right, Fig. 12 (a), (b) and (c) show the potential volume gain per tile from existing connections, new services



(a) Existing connections



(b) New connections

FIGURE 10. CDF of estimated data volume gain.

and the sum of both, respectively. It is observed that the potential volume gained by improving already established connections, $M_{\Delta V_{opt}}(x,y)$, is more distributed across the map. In contrast, the gain from new services, $M_{\Delta V_{new}}(x,y)$, is more concentrated in specific areas. This is due to the fact many rings do not have non-last TTI connections. In both maps, transparent tiles show areas with zero gain, where no connections were established. Most of these tiles are in unpopulated areas, out of the targeted coverage region. Table 7 confirms these findings by presenting some statistics of the three spatial distributions. As shown in the figures and the table, the largest volume gains per tile come from new connections. Likewise, the overall gain map can be used to detect areas already performing at optimal conditions (i.e., low $M_{\Delta V}(x,y)$) and others with bad radio link conditions and many users (i.e., high $M_{\Delta V}(x,y)$ value).

2) SMALL CELL SELECTION

Finally, the decision of where to locate the new SCs is made. To this end, the potential data volume gains per tile are aggregated to compute the total traffic gain for the different candidate SC locations, $G_{sc}(x,y)$. Fig. 13 presents the 100 largest total gain values, ordered from highest to lowest.

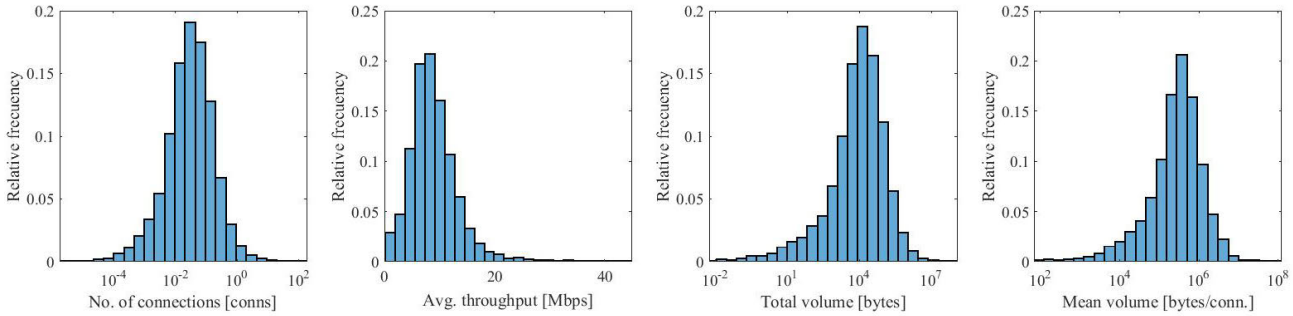
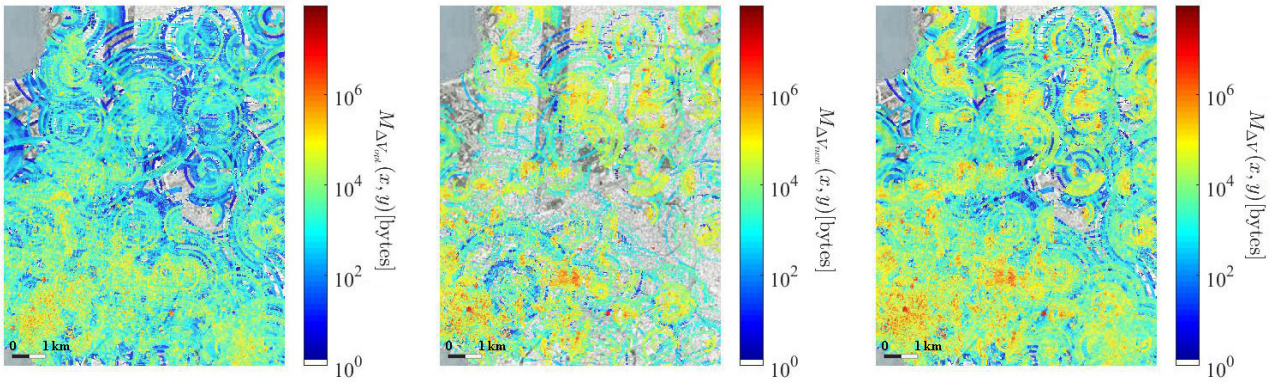


FIGURE 11. Statistical distribution of indicators computed per tile.



(a) Existing connections (b) New services (c) Existing connections+new services

FIGURE 12. Spatial distribution for potential data volume improvement, $M_{\Delta V}(x, y)$.

TABLE 8. Examples of new sites where methods perform differently.

Case	Description	Example
Case 1	Site for blackspot suggested by both TSM and TM/SM, but in a slightly different position.	Fig. 17(a)
Case 2	Site for blackspot suggested only by TSM.	Fig. 17(b)
Case 3	Site for hotspot suggested by both TSM and TM/SM, but in a slightly different position.	Fig. 19(a)
Case 4	Site for hotspot suggested only by TSM.	Fig. 19(b)
Case 5	Site suggested only by TM/SM.	Fig. 19(c)

For a more detailed analysis, gains are broken down into that coming from existing and from new connections. It is observed that, in most sites, traffic gains come from new connections.

To show the ability of the method to detect coverage issues, Fig. 14 plots two examples of selected SC locations (dot) and their ideal coverage areas (circle of 50-meter radius) over a coverage map extracted from OpenSignal platform [20]. This platform collects geolocated signal level measurements anonymously from mobile users subscribed to this initiative. In the figure, it is observed that the proposed SC sites would cover areas reported as of weak radio signal level by OpenSignal (red areas in the figure).

Finally, the analysis is focused on the indoor gain ratio indicator, $R_{in}(x_{sc}, y_{sc})$, reflecting how much of the data volume gain from a SC comes from indoor tiles. Fig. 15 shows the histogram of $R_{in}(x_{sc}, y_{sc})$ for the best 100 candidate SCs. It is observed that the indoor gain ratio of the best sites tends to be above 0.5, showing that the traffic gain is mainly originated by indoor users. Specifically, 73 of the 100 best sites have indoor gain ratio greater than 0.5 and should be located indoors.

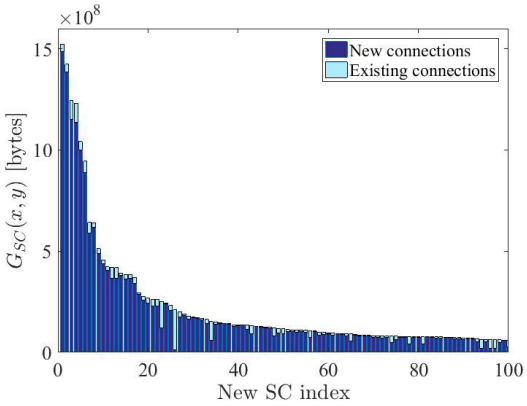
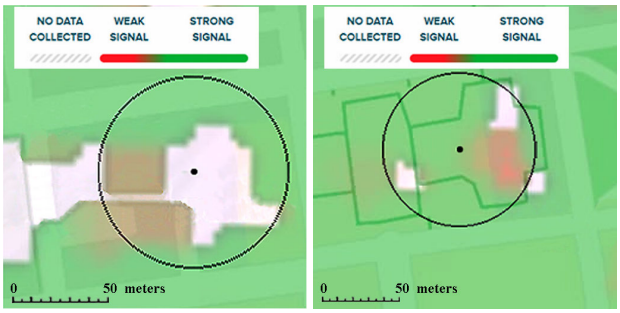


FIGURE 13. Data volume increase for the best 100 small cell locations.



(a) Location 1 (b) Location 2

FIGURE 14. Example of selected SC sites and received signal level (OpenSignal).

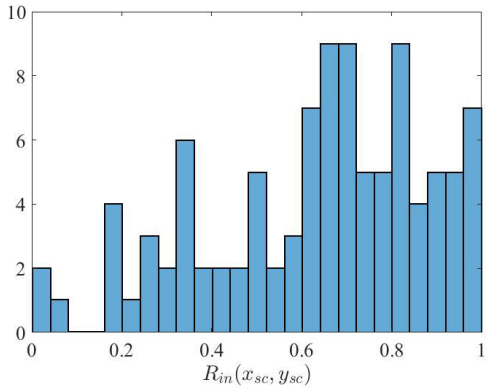


FIGURE 15. Histogram of indoor gain ratio of the best 100 candidate SC sites.

The proposed method that combines trace and social network data (TSM) is compared with legacy methods that only use traces (TM) or social data (SM). Fig. 16 shows the CDF of the data volume gains obtained from the best 100 new SCs suggested by the three methods, evaluated with the potential data volume gain map of TSM. As expected, TSM obtains larger data volume gains per site. Overall, TSM achieves a total data volume gain in the network of 27.58 GB, TM of 25.40 GB (8% less) and SM only of 3.22 GB (88% less). This result points out that the solution obtained by SM greatly differs from that of TSM and TM. A detailed analysis shows that the difference between TM and TSM is not large because many TA rings have few posts due to the limited size of the social network dataset, causing traffic in these rings to be evenly distributed (per context) in TSM, as in TM. It is expected that larger differences would be observed with a larger post dataset.

A more detailed analysis of SC locations on a map shows important differences among methods. To this end, the output of the methods is compared to find cases where the proposed TSM method perform differently from legacy methods, TM and SM. For convenience, the new SCs are divided in two groups, depending on whether their aim is to cover areas with poor coverage (i.e., black spots) or high traffic demand (i.e., hot spots). Table 8 summarizes the identified five cases, presented next. In both groups, several cases are analyzed where: a) the problematic spot is detected by several methods,

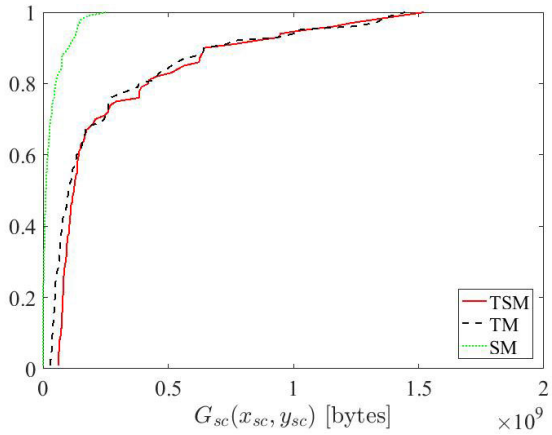


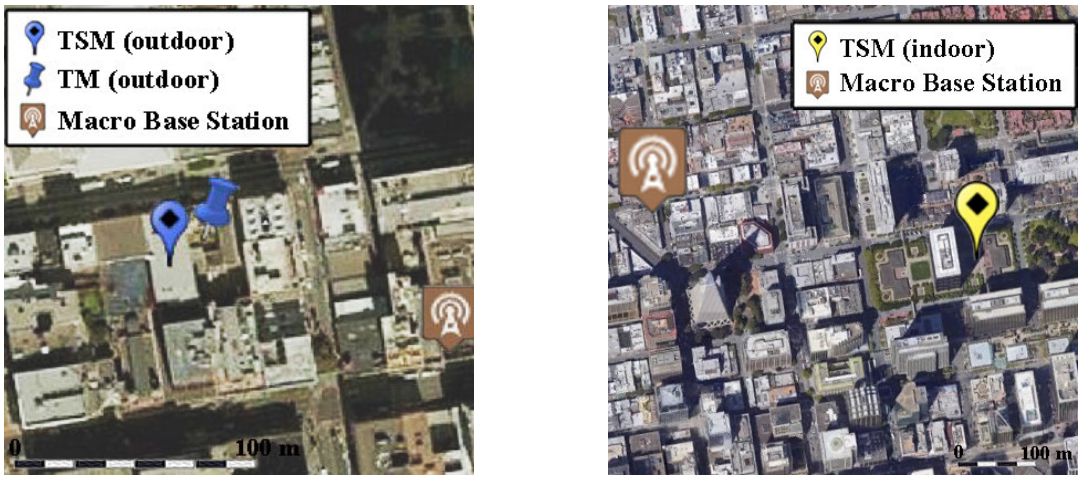
FIGURE 16. Distribution of data volume gain of the best 100 candidate SC sites.

but in slightly different locations, or b) the problematic spot is only detected by one of the methods.

Fig. 17(a) shows the first case of an outdoor blackspot detected by TSM and TM. For a complete picture, the figure depicts the locations of the SCs suggested by TSM and TM, the nearby macro base stations and the post messages. Not shown is the fact that both SC locations cover an area not in line of sight with the macro base station due to a tall building (and, hence, the blackspot). It is observed that the SC location of TSM is shifted to the right, following the post distribution. In spite of this displacement, the total data volume gain estimated by TSM and TM is almost the same (1437.4 MB and 1423.7 MB, respectively). In both cases, most of the gain comes from outdoor tiles.

Fig. 17(b) shows the second case of an indoor blackspot only detected by TSM. A preliminary analysis shows that the building where the SC is located is a underground parking garage, which is in non-line-of-sight conditions with the nearby macro base station due to a skyscraper. To understand why TM does not suggest a site for this area, Fig. 18 shows the number of connections per tile, $M_k(x, y)$, and the associated traffic gain, $G_{sc}(x, y)$, used by TSM and TM in the area of the SC suggested by TSM. On the left, it is observed that the connection density in TM is low and regular, which is the result of distributing connections in TA rings of the macrocell evenly in space (per context). In contrast, connection density in TSM is large and irregular, due to the concentration of posts inside the car park. This difference justifies the reason for the large deviations in expected traffic gains. Moreover, note that, despite the large number of posts in the area, SM does not suggest a new SC due to its inability to detect coverage problems.

Fig. 19(a) shows the third case of an indoor hotspot detected by TSM and TM. In this case, TSM locates the new SC in the corner of a skyscraper hosting an important company thanks to the post messages, whereas TM locates the new SC in a smaller (and possibly less populated) nearby building. As a result, the total gain achieved by the new SC (estimated with traffic map of TSM) is 125.81 MB for the



(a) First case. (b) Second case.

FIGURE 17. Small cell proposals for black spot area.

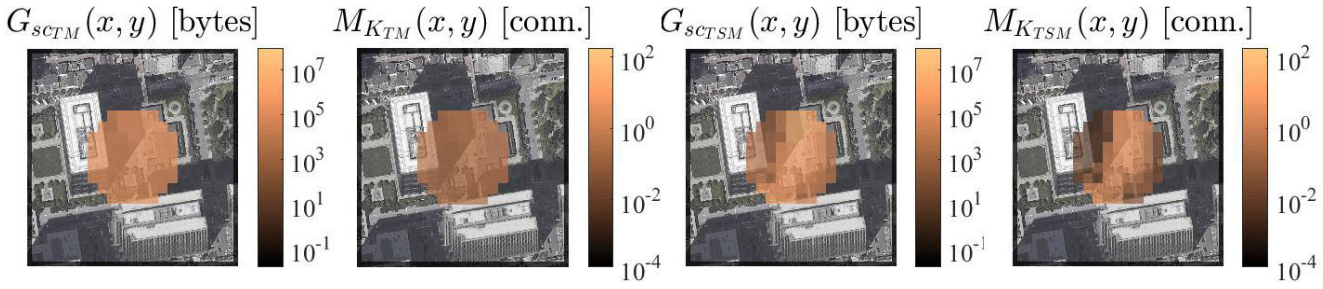


FIGURE 18. Detailed analysis of TM and TSM (second case).

(a) Third case. (b) Fourth case. (c) Fifth case.

FIGURE 19. Small cell proposals for hotspot area.

position suggested by TSM and only 67.59 MB for that of TM. In this case, SM fails to include the site in the list of 100 best SC locations because the number of posts in the area is not high enough.

Fig. 19(b) illustrates the case of an indoor hotspot detected only by TSM and SM. It is not seen in the figure that the SC suggested by TSM is inside a shopping center, which emphasizes the importance of considering geolocated posts. Fig. 20 presents the same detailed analysis, showing that TM fails to detect the hotspot due to its difficulty to geolocate connections more precisely than in rings. In this case, SM detects two points of user concentration by searching for peaks in the post spatial distribution.

Finally, Fig. 19(c) depicts a case where SM detects a hotspot that is not detected by TSM and TM. This case can be explained by a peak of posts in a region where mobile users have proper coverage and enough available radio resources.

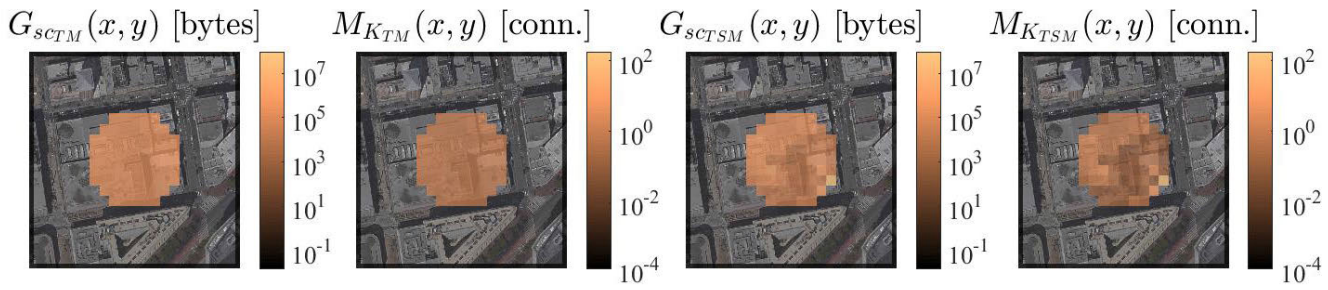


FIGURE 20. Detailed analysis of TM and TSM (fourth case).

D. COMPUTATIONAL ISSUES

The proposed method needs some previous work (collection/pre-processing of traces and construction of land use/post map) before the procedure in Fig. 1 can be launched. The execution time of pre-processing traces grows linear with the number of connections and data fields, while the construction of maps grows linear with the number of tiles in the map. Once input data is available, the computational complexity of the method is given by the algorithm for building the spatial connection distribution. The algorithm distributes N_i indicators from $N_{conn}(r)$ connections originated in the area covered by a ring r of the N_r rings in the scenario. Thus, the worst-case time complexity is $\mathcal{O}(N_r * N_{conn} * N_i)$.

Trace processing is done by complex event processing with Esper routines [59]. Land use are processed with Matlab, post data are obtained with the Streaming API using Java through the library Twitter4j and the Flickr API using Python [25], [26], [60] and processed with Matlab. The proposed method is implemented with the Statistics and Machine Learning Toolbox and Image Processing Toolbox in Matlab. All processes are executed in a server with a 2.4-GHz octa-core processor and 64 GB of RAM. The time required for decoding connection traces (400 cells, 2 hours of traces, 166,561 connections) is 282 seconds. The time to build the land use and post maps (125 km^2, 1,222,787 tiles) are 50,444 and 57,743 seconds. Finally, the time to detect the 100 best candidate sites with the above dataset is 238 seconds, 67% of which is spent in the construction of the spatial connection distribution.

V. CONCLUSION

Small-cell site selection is currently a labor-intensive process. In this paper, an automatic context-aware data-driven method has been proposed to precisely detect small areas with coverage or capacity problems in a mobile network based on the performance of connections. The core of the method is the positioning of connections based on the indoor probability of each connection and the distribution of geolocated posts from social networks. The method has been tested with a large trace dataset from a live Long Term Evolution network and a database of geotagged posts from Twitter and Flickr.

Results have shown that problems detected in the network by combining connection data and geotagged posts are consistent with their context, i.e., sites detected due to poor coverage present bad propagation conditions from the serving macrocell, while spots with capacity problems are located in very populated places (e.g., museums, schools, shopping centers, etc.). Likewise, the indoor/outdoor distinction for the new small cell is coherent, i.e., sites tagged as indoor are located in indoor locations, whereas spots covering open areas in the city are classified as outdoor. Moreover, comparison with legacy approaches have shown important differences in the sites selected.

A key component in the proposed method is regression curves modeling the impact of user throughput on traffic volume and service mix. Figures have shown that regression accuracy can still be improved. For this purpose, sophisticated regression models considering more predictors can be derived with machine learning techniques, provided that a large and diverse measurement dataset is available.

The low computational complexity of the method allows an easy integration in radio planning tools. By combining different data sources, the method can make the most of the latest big-data empowered network management systems.

REFERENCES

[1] Ericsson. (2018). *Ericsson Mobility Report June 2018*. pp. 3–13. Accessed: May 24, 2019. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf/

[2] 5G PPP Architecture Working Group *et al.*, "5G empowering vertical industries," *5G-PPP Roadmap*, no. 9, Feb. 2016. Accessed: May 24, 2019. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf/

[3] J. Ramiro and K. Hamied, *Self-Organizing Networks: Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*. Hoboken, NJ, USA: Wiley, 2011.

[4] S. Barakovic and L. Skorin-Kapov, "Survey and challenges of QoE management issues in wireless networks," *J. Comput. Netw. Commun.*, vol. 2013, pp. 1–28, Mar. 2013.

[5] A. Banerjee, "Revolutionizing CEM with subscriber-centric network operations and QoE strategy," Heavy Reading, White Paper, 2014. Accessed: May 24, 2019. [Online]. Available: http://www.accantosystems.com/wp-content/uploads/2018/05/Heavy-Reading-Accanto.pdf/

[6] P. Le Callet *et al.*, "Qualinet white paper on definitions of quality of experience, version 1.1," in *Proc. Eur. Netw. Qual. Exper. Multimedia Syst. Services (COST Action IC)*, vol. 3, 2012, pp. 1–23.

[7] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Comput.*, vol. 6, no. 3, pp. 30–38, Jul. 2007.

[8] P. Makris, D. N. Skoutas, and C. Skianis, "A survey on context-aware mobile and wireless networking: On networking and computing environments' integration," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 362–386, 1st Quart., 2013.

[9] M. Keally, G. Zhou, G. Xing, J. Wu, and A. Pyles, "PBN: Towards practical activity recognition using smartphone-based body sensor networks," in *Proc. 9th ACM Conf. Embedded Networked Sensor Syst. (SenSys)*, 2011, pp. 246–259.

[10] N. Baldo, L. Giupponi, and J. Mangues, "Big data empowered self organized networks," in *Proc. 20th Eur. Wireless Conf.*, May 2014, pp. 1–8.

[11] W. Wang, Q. Chang, Q. Li, Z. Shi, and W. Chen, "Indoor-outdoor detection using a smart phone sensor," *Sensors*, vol. 16, no. 10, p. 1563, Sep. 2016.

[12] J. L. Bejarano-Luque, M. Toril, M. Fernandez-Navarro, R. Acedo-Hernandez, and S. Luna-Ramirez, "A data-driven algorithm for Indoor/Outdoor detection based on connection traces in a LTE network," *IEEE Access*, vol. 7, pp. 65877–65888, 2019.

[13] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov. 2014.

[14] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan. 2016.

[15] S. Fortes, A. Aguilar-García, R. Barco, F. Barba, J. Fernández-luque, and A. Fernández-Durán, "Management architecture for location-aware self-organizing LTE/LTE–A small cell networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 294–302, Jan. 2015.

[16] 5G PPP Architecture Working Group *et al.*, "View on 5G architecture," *5G-PPP White Papers*, no. 1, 2016. Accessed: Jun. 1, 2020. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf/

[17] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[18] W. A. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, and B. Sebire, "Minimization of drive tests solution in 3GPP," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 28–36, Jun. 2012.

[19] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Automated coverage hole detection for cellular networks using radio environment maps," in *Proc. 11th Int. Symp. Workshops Modeling Optim. Mobile, Ad Hoc Wireless Netw. (WiOpt)*, May 2013, pp. 35–40.

[20] B. Gill, S. Khanifar, J. Robinson, and S. Westwood. (2010). *OpenSignal*. Accessed: Jun. 20, 2019. [Online]. Available: https://www.opensignal.com/

[21] F. Malandrino, C.-F. Chiasserini, and S. Kirkpatrick, "Cellular network traces towards 5G: Usage, analysis and generation," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 529–542, Mar. 2018.

[22] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, "Estimating mobile traffic demand using Twitter," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 380–383, Aug. 2016.

[23] S. Fortes, D. Palacios, I. Serrano, and R. Barco, "Applying social event data for the management of cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 36–43, Nov. 2018.

[24] W. Guo and J. Zhang, "Uncovering wireless blackspots using Twitter data," *Electron. Lett.*, vol. 53, no. 12, pp. 814–816, Jun. 2017.

[25] (2006). *Twitter API Documentation*. Accessed: Apr. 2, 2019. [Online]. Available: https://dev.twitter.com/docs

[26] (2004). *Flickr API Documentation*. Accessed: Apr. 2, 2019. [Online]. Available: https://www.flickr.com/services/api/

[27] H. R. Anderson and J. P. McGeehan, "Optimizing microcell base station locations using simulated annealing techniques," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Jun. 1994, pp. 858–862.

[28] C. Y. Lee and H. G. Kang, "Cell planning with capacity expansion in mobile communications: A tabu search approach," *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1678–1691, 2000.

[29] A. J. Nebro, E. Alba, G. Molina, F. Chicano, F. Luna, and J. J. Durillo, "Optimal antenna placement using a new multi-objective CHC algorithm," in *Proc. 9th Annu. Conf. Genetic Evol. Comput. (GECCO)*, 2007, pp. 876–883.

[30] L. Raisanen and R. M. Whitaker, "Comparison and evaluation of multiple objective genetic algorithms for the antenna placement problem," *Mobile Netw. Appl.*, vol. 10, nos. 1–2, pp. 79–88, Feb. 2005.

[31] M. J. Nawrocki and T. W. Wieckowski, "Optimal site and antenna location for UMTS output results of 3G network simulation software," in *Proc. 14th Int. Conf. Microw., Radar Wireless Commun. (MIKON)*, vol. 3, May 2002, pp. 890–893.

[32] J. Niemelä and J. Lempiäinen, "Impact of base station locations and antenna orientations on UMTS radio network capacity and coverage evolution," in *Proc. IEEE Int. Symp. Wireless Pers. Multimedia Commun.*, Yokosuka, Japan, Oct. 2003.

[33] E. Amaldi, A. Capone, and F. Malucelli, "Planning umts base station location: Optimization models with power control and algorithms," *IEEE Trans. Wireless Commun.*, vol. 2, no. 5, pp. 939–952, Sep. 2003.

[34] L. Nagy and L. Farkas, "Indoor base station location optimization using genetic algorithms," in *Proc. 11th IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, vol. 2, Sep. 2000, pp. 843–846.

[35] Z. Ji, T. K. Sarkar, and B.-H. Li, "Methods for optimizing the location of base stations for indoor wireless communications," *IEEE Trans. Antennas Propag.*, vol. 50, no. 10, pp. 1481–1483, Oct. 2002.

[36] L. Nagy, "Global optimization of indoor radio coverage," *Automatika*, vol. 53, no. 1, pp. 69–79, Jan. 2012.

[37] S. Rodd, A. Prof, and A. H. Kulkarni, "Optimization algorithms for access point deployment in wireless networks," *J. Comput. Appl.*, vol. 2, no. 1, p. 2, 2009.

[38] J. K. L. Wong, A. J. Mason, M. J. Neve, and K. W. Sowerby, "Base station placement in indoor wireless systems using binary integer programming," *IEE Proc.-Commun.*, vol. 153, no. 5, pp. 771–778, Oct. 2006.

[39] Y. Ngadiman, Y. H. Chew, and B. S. Yeo, "A new approach for finding optimal base stations configuration for CDMA systems jointly with uplink and downlink constraints," in *Proc. IEEE 16th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2005, pp. 2751–2755.

[40] L. K. Pujji, K. W. Sowerby, and M. J. Neve, "A new algorithm for efficient optimisation of base station placement in indoor wireless communication systems," in *Proc. 7th Annu. Commun. Netw. Services Res. Conf.*, May 2009, pp. 425–427.

[41] S. Wang, W. Guo, and T. O'Farrell, "Optimising femtocell placement in an interference limited network: Theory and simulation," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–6.

[42] M. Talau, E. C. G. Wille, and H. S. Lopes, "Solving the base station placement problem by means of swarm intelligence," in *Proc. IEEE Symp. Comput. Intell. Commun. Syst. Netw. (CIComms)*, Apr. 2013, pp. 39–44.

[43] L. K. Pujji, K. W. Sowerby, and M. J. Neve, "Development of a hybrid algorithm for efficient optimisation of base station placement for indoor wireless communication systems," *Wireless Pers. Commun.*, vol. 69, no. 1, pp. 471–486, Mar. 2013.

[44] J. Liu, Q. Chen, and H. D. Sherali, "Algorithm design for femtocell base station placement in commercial building environments," in *Proc. IEEE Infocom*, Mar. 2012, pp. 2951–2955.

[45] J. M. Ruiz Avilés, M. Toril, and S. Luna-Ramírez, "A femtocell location strategy for improving adaptive traffic sharing in heterogeneous LTE networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 38, Dec. 2015.

[46] C. Sharma, Y. F. Wong, W.-S. Soh, and W.-C. Wong, "Access point placement for fingerprint-based localization," in *Proc. IEEE Int. Conf. Commun. Syst.*, Nov. 2010, pp. 238–243.

[47] K. Farkas, A. Huszák, and G. Gódor, "Optimization of Wi-Fi access point placement for indoor localization," *J. IIT (Inform. IT Today)*, vol. 1, no. 1, pp. 28–33, 2013.

[48] F. Al-Turjman, E. Ever, and H. Zahmatkesh, "Small cells in the forthcoming 5G/IoT: Traffic modelling and deployment overview," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 28–65, 1st Quart., 2019.

[49] H. Klessig, H. Kuntzschmann, L. Scheuvens, B. Almeroth, P. Schulz, and G. Fettweis, "Twitter as a source for spatial traffic information in big data-enabled self-organizing networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–5.

[50] A. Sánchez, R. Acedo-Hernández, M. Toril, S. Luna-Ramírez, and C. Úbeda, "A trace data-based approach for an accurate estimation of precise utilization maps in LTE," *Mobile Inf. Syst.*, vol. 2017, pp. 1–10, May 2017.

[51] M. Masse, *REST API Design Rulebook: Designing Consistent Restful Web Service Interfaces*. Newton, MA, USA: O'Reilly Media, Inc, 2011.

[52] S. Coast. (2004). *OpenStreetMap*. Accessed: Nov. 4, 2018. [Online]. Available: https://www.openstreetmap.org/

[53] V. Buenestado, J. M. Ruiz-Aviles, M. Toril, S. Luna-Ramirez, and A. Mendo, "Analysis of throughput performance statistics for benchmarking LTE networks," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1607–1610, Sep. 2014.

[54] *Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E UTRAN): Definitions, Version 9.1.0 Release 9*, document 32.450, 3GPP, Jun. 2010.

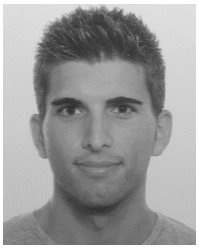
[55] M. A. Oliver and R. Webster, "Kriging: A method of interpolation for geographical information systems," *Int. J. Geographical Inf. Syst.*, vol. 4, no. 3, pp. 313–332, Jul. 1990.

[56] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, and L. Nan, "Trends in small cell enhancements in LTE advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 98–105, Feb. 2013.
[57] *Use Case Characterization, KPIs and Preferred Suitable Frequency Ranges for Future 5G Systems Between 6 GHz And 100 GHz, mmMAGIC Deliverable 1.1 Version 1*, document ICT-671650, Nov. 2015.
[58] J. L. Bejarano-Luque and A. J. Garcia. *Codes for Post Collection*. Accessed: May 22, 2020. [Online]. Available: https://mobilenet.uma.es/index.php/resources/
[59] EsperTech Inc. (2006). *Esper: Language, Compiler and Runtime forbreak Complex Event Processing (CEP)*. Accessed: Jan. 14, 2020. [Online]. Available: http://www.espertech.com/esper/
[60] Y. Yamamoto. (2007). *Twitter4j Library*. Accessed: Apr. 2, 2019. [Online]. Available: http://twitter4j.org/en/

JUAN L. BEJARANO-LUQUE received the B.S. degree in telecommunications engineering and the M.S. degree in acoustic engineering from the University of Málaga, Málaga, Spain, in 2015 and 2016, respectively, where he is currently pursuing the Ph.D. degree in telecommunications engineering. His research interests include optimization of radio resource management for mobile networks, location-based services and management, and data analytics.

MATÍAS TORIL received the M.S. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Spain, in 1995 and 2007, respectively. Since 1997, he has been a Lecturer with the Communications Engineering Department, University of Málaga, where he is currently a Full Professor. He has coauthored more than 100 publications in leading conferences and journals and three patents owned by Nokia Corporation. His current research interests include self-organizing networks, radio resource management, and data analytics.

MARIANO FERNÁNDEZ-NAVARRO received the M.S. degree in telecommunication engineering from the Polytechnic University of Madrid, in 1988, and the Ph.D. degree from the University of Málaga, in 1999. He is on the Staff with the Communications Engineering Department, University of Málaga since 1992, after three years as a Design Engineer with Fujitsu Spain S. A. His research interests include optimization of radio resource management for mobile networks and location-based services and management.



ANTONIO J. GARCÍA received the M.S. degree in telecommunication engineering from the University of Málaga, Spain, in 2014. He is currently pursuing the Ph.D. degree in telecommunications engineering in a Collaborative Project with Ericsson. Since 2014, he has been with the Communications Engineering Department, University of Málaga. His research interest includes planning and optimization of mobile radio access networks based on users experience.

SALVADOR LUNA-RAMÍREZ (Member, IEEE) received the M.S. degree in telecommunication engineering and the Ph.D. degree from the University of Málaga, Spain, in 2000 and 2010, respectively. He has been a Lecturer with the Communications Engineering Department, University of Málaga, since 2000, where he is currently a Full Professor. His research interests include self-optimization of mobile radio access networks and management of radio resources, in addition to research and collaboration with companies in the field of acoustic engineering.