This document is a preprint of: H. Klessig, V. Suryaprakash, O. Blume, A. Fehske and G. Fettweis, "A Framework Enabling Spatial Analysis of Mobile Traffic Hot Spots," in IEEE Wireless Communications Letters (WCL), Oct 2014. DOI:10.1109/LWC.2014.2349520

A Framework Enabling Spatial Analysis of Mobile Traffic Hot Spots

Henrik Klessig, *Member, IEEE*, Vinay Suryaprakash, *Member, IEEE*, Oliver Blume, *Member, IEEE*, Albrecht Fehske, *Member, IEEE*, and Gerhard Fettweis, *Fellow, IEEE*

***Abstract*—An enormous increase in data traffic demanded by mobile users calls for efficient deployment strategies such as multi-layer heterogeneous networks. However, placing small cells at the desired locations to offload as much traffic as possible from overlaying macro cells is a crucial task. In this regard, geo-location and user equipment positioning techniques help obtain spatial distributions of user locations and their respective traffic volumes. In this paper, we provide a tool capable of reducing errors that stem from spatial discretization of traffic data and that can autonomously detect hot spots given a certain threshold. Based on geo-located traffic in a 3G network in a dense urban city, we find that traffic in the area is approximately log-normally distributed and that the size of traffic hot spots are approximately Weibull distributed. Based on our statistical findings, we observe that utilizing 4 small cells per km^2 covering 3.2% of the total area and around 34% of the total traffic volume is a very meaningful deployment strategy; however, deploying more small cells in larger hot zones becomes increasingly costly in terms of the ratio of area covered and traffic demand serviced.**

***Index Terms*—Mobile data traffic, hot spots, traffic analysis, network planning, network deployment, small cells.**

I. Introduction

SPATIAL traffic modeling and analysis has recently garnered momentum due to two reasons: The first is because of an enormous increase in mobile data traffic [1], which has resulted in effective as well as efficient deployment strategies (such as multi-layer heterogeneous networks or ultra-dense small cell networks, see e.g., [2], [3]) requiring precise knowledge of spatial traffic patterns to prevent over- and under-provisioning of capacity. The second is due to the ever improving resolution of the geo-location and User Equipment (UE) positioning techniques (which provide the spatial distribution of mobile phone users and their data volumes) offered by base station vendors and network optimization experts, [4]. Consequently, accurate data provided by the network can be used more frequently in the future, which helps take steps towards improved network planning and a more definite network performance evaluation.

Plenty of work, in papers such as [5] and [6], has been carried out on the analysis of temporal long-term and short-term mobile traffic fluctuations, for example, through measurements of data volumes transferred by base stations. Recently, spatio-temporal investigations have also been carried out. For instance, based on mobile phone data, the authors in [7] investigate the number of hot spots (locations where mobile phone users congregate) as a function of the overall city population and observed spatial and temporal stationarity of such hot spots, which form the *hearts of cities*. In the most recent works [8] and [9], the authors analyze spatial traffic distributions based on EDGE/GPRS traffic volumes per cell recorded by the base stations. They distributed the data collected evenly over the Voronoi cells formed by the locations of the base stations and found that traffic is log-normally distributed over the entire area and Weibull distributed across the cells. Though these results are very promising with respect to the ease of generating traffic maps for simulation, the spatial resolution of such methods is, in general, limited by the cell sizes. When compared to a dense urban environment, results may be less convincing for regions ranging from urban to rural in which base stations are deployed more sparsely. Furthermore, the spatial distribution within the cells remains obscured, which is a fundamental problem if the distribution of traffic on the spatial scale of small cells is of interest. Moreover, EDGE/GPRS traffic may be less representative as the traffic volumes observed might be limited by base station capacities rather than being determined by user demands.

The main drawback of most of the existing work is the coarse spatial resolution that is inherent to distributing accumulated traffic data measured in the base station evenly within the cells' areas. To the best knowledge of the authors, detailed spatial analyses of traffic hot spots based on geo-location and UE positioning techniques do not exist. In this paper, we analyze mobile data traffic transmitted to the users in the downlink of a 3G network in a dense urban European city. The traffic is localized with the aid of common geo-location techniques. Thereby, the spatial resolution is not dependent on the deployment density and is comparable to the size of the average pico cell. The contributions of the paper are three-fold:

1) We develop a methodology that allows generating mobile data traffic maps from data sets provided by a fully operational cellular network and is capable of reducing errors that stem from data collection.
2) We apply an image processing technique to autonomously detect *hot spots* and *hot zones*, i.e., locations where the density of data volumes transmitted is high compared to the rest of the area, based on specific values of thresholds. Though the definition of the thresholds can be customized, we provide a recommendation for the best method of defining thresholds as well.
3) We analyze traffic maps that result from the data processing steps and an example data set from a 3G network[1] and mainly focus on the statistical traffic distribution, the density and size of hot spots, as well as, the traffic therein.

Manuscript received June 16, 2014; accepted August 5, 2014. Date of publication August 19, 2014; date of current version October 9, 2014. This work was supported in part by the GreenTouch Consortium. The associate editor coordinating the review of this paper and approving it for publication was H. Li.

H. Klessig, V. Suryaprakash, and G. Fettweis are with the University of Technology Dresden, Vodafone Chair Mobile Communications Systems, Dresden, Germany.

O. Blume is with Alcatel-Lucent Bell Labs., 70435 Stuttgart, Germany.

A. Fehske is with Airrays GmbH, Dresden, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LWC.2014.2349520

[1]The data processing steps can be applied to any other wireless network technology that is capable of collecting call traces.

2162-2337 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This document is a preprint of: H. Klessig, V. Suryaprakash, O. Blume, A. Fehske and G. Fettweis, "A Framework Enabling Spatial Analysis of Mobile Traffic Hot Spots," in IEEE Wireless Communications Letters (WCL), Oct 2014. DOI:10.1109/LWC.2014.2349520

II. Traffic Data Description and Processing

We use a data set of call traces of a 3G mobile network located in an urban northern European city to illustrate the data processing and analysis chain, and to provide statistical findings. This call trace data, which was collected by the base stations in December 2013 over a period of four consecutive working days, contains spatial and time resolved measurements of the data volume demanded by the users in the downlink. We restrict the evaluation of the data set to a *dense* urban sub-region of size 16.78 km^2, because, more often than not, hot spots (and their stationary behavior) have also been observed in other urban areas in Europe [7]. Furthermore, we restrict our observations to a total time interval of the 15 busiest hours because, in times of low traffic, isolated events can significantly distort the distribution. This subset of the data is processed to identify stationary locations of higher data traffic, the relative fraction of hot spot traffic, and the size of the hot spots. Furthermore, we normalize the data in the area such that it is given as a binned *traffic demand density* in Mbps/km^2. We observe an average traffic density of 5.09 Mbps/km^2.

A. Measurement Techniques and Spatial Resolution

The location data is determined based on a combination of various geo-location techniques, namely signal-delay based estimates like Observed Time Difference of Arrival (OTDOA), estimation based on pathloss and network configuration (e.g., antenna azimuth), and techniques based on Cell-ID's.[2] There exist two types of errors that feature in the data provided.

1) Location Errors: These errors, also known as geo-location errors, are usually caused by fluctuations in receive powers and signal delays, more often than not, driven by user mobility or changing environmental conditions. Assuming the localization errors of the UE positions to be i.i.d., and given the distribution of the localization error, one could eliminate the error by deconvolving the raw data with the error distribution. Since the raw data may be noisy (due to a limited number of samples), we recommend using deconvolution techniques like the Lucy-Richardson deconvolution, see e.g., [10] and [11]. Since the distribution of the location error of the data given in this paper is unknown to the authors, this step is left out of the processing chain; however, it is important to note that the scale of the geo-location error might be the same as that of the discretization error due to binning, which is described subsequently.

2) Discretization Errors Due to Binning: The conceivably complex combination of geo-location techniques makes a certain "maximum" spatial resolution necessary, such that the raw data obtained is summed up over a bin and this traffic value is assigned to the mid-point of the bin (spatial discretization). Usually, all proprietary network analysis tools exhibit such a maximum resolution. Summing data over the bin turns out to be quite a destructive process because it removes valuable information about the actual distribution of the traffic within the bins. Moreover, hot spot analysis of the raw data (without the steps used in this work) would result in discretized hot spot sizes (in this case) in steps of 3×3 arc seconds, which, in central Europe, corresponds to an area of approx. 8000 m^2. Since this resolution is rather coarse, we need to apply certain interpolation techniques, which are explained in the following.

[2]All of the methods of localization listed are widely used; however, the details of their inner working and the manner in which they are combined is done proprietarily and cannot be provided here in greater detail.

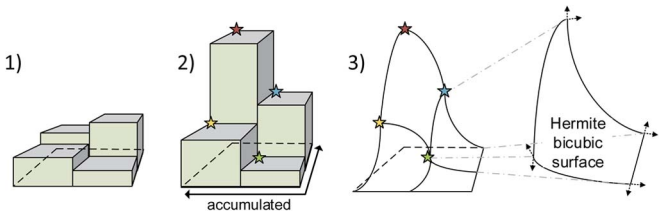

Fig. 1. (1) Binned traffic density as a p.m.f. (with 4 bins); (2) cumulative sum; (3) cumulative traffic after Hermite bicubic interpolation.

B. Eliminating Discretization Errors by Hermite Interpolation

To increase the spatial resolution of the data by considering the *difference* between adjacent bin values, we could interpret the binned traffic as a histogram and apply interpolation techniques to it. However, the interpolation has to take place constrained by the facts that, the amount of total traffic in each bin remains the same and that all resulting values are non-negative. Hermite bicubic interpolation of the cumulative sum (c.d.f.) of the histogram (p.m.f.) is a method that can provide just such a result. The idea is sketched in Fig. 1. After interpolating the cumulative traffic (see step "3)" in Fig. 1) and taking its derivative, we obtain a smoothed version of the traffic density with the constraints given above. The traffic map now has a much finer spatial granularity given as *pixels* whose size are much smaller than that of the bins of the raw data. For mathematical and implementation details, we refer the reader to [12].

C. Identifying Hot Spots and Hot Zones

1) Definition of Hot Spots and Hot Zones: We define a traffic *hot spot* to be a region in the plane in which, the traffic (demand) density exceeds a certain threshold and can usually be covered by one small cell. More importantly, we do not define them as locations in which the number of users or devices is high, since network performance is mainly driven by the amount of data requested rather than the number of connections. Furthermore, we define a *hot zone* to be an extended and arbitrarily shaped region, which is too large to be served by a single small cell, the traffic exceeds the average traffic value, and contains a considerable share of the overall traffic. Hot zones may contain one or multiple hot spots.

2) Thresholding: Thresholding is a process carried out to determine locations (or pixels) which belong to hot spots (or hot zones). Considering a fixed (arbitrary) threshold for classifying hot spots is meaningless when traffic grows over time and network capacity is upgraded continuously. Hence, we require the hot spots detected to remain unchanged during the course of a time. Since the mean (viz. the expectation) scales in proportion to a scaling in the overall traffic by definition, the hot spots (and their sizes) are robust to portended increases in traffic over months/years. Hence, choosing multiples of the mean is a requirement to achieve this goal. In contrast, considering percentiles of the traffic distribution as thresholds defines the percentage of the area containing hot spots *a priori*. This amounts to pre-selecting a preferred business model with

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This document is a preprint of: H. Klessig, V. Suryaprakash, O. Blume, A. Fehske and G. Fettweis, "A Framework Enabling Spatial Analysis of Mobile Traffic Hot Spots," in IEEE Wireless Communications Letters (WCL), Oct 2014. DOI:10.1109/LWC.2014.2349520

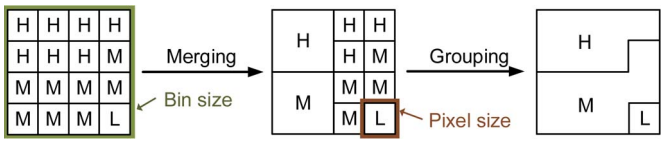

Fig. 2. (Left) Traffic density map with categories *high* (H, e.g., hot spots), *medium* (M, e.g., hot zones), and *low* (L) traffic; (center) after merging; (right) after merging and grouping.

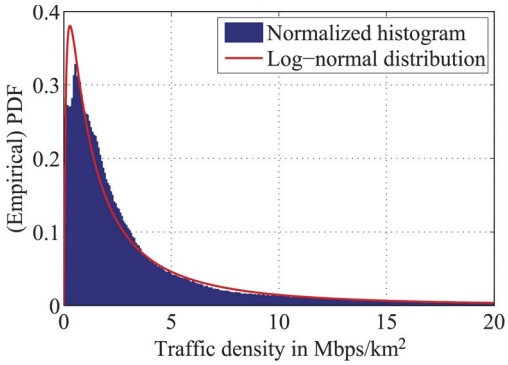

Fig. 3. Normalized histogram of the spatial traffic distribution with log-normal fit.

respect to small cell deployment, which is undesirable for a fair assessment of hot spots.

3) Merging and Grouping of High Traffic Regions: Merging is done to coalesce neighboring pixels having a similar traffic density (defined by the two thresholds characterizing hot spots and hot zones) to form larger areas. Grouping is carried out to form larger irregular regions by grouping coalesced pixels (see Fig. 2 for an illustration). These two processing steps, merging and grouping, are methods that are used in pattern recognition and object detection, and are described in [13]. It is important to note that these steps are required to ensure that our algorithm to detect hot spots can distinguish between individual hot spots and analyze their characteristics (such as sizes, shapes, etc.) autonomously. They are also independent of the trace collection processes described in Section II-A.

III. Analysis of Mobile Traffic Hot Spots

Once the raw data is processed, we carry out statistical analysis to determine hot spot related parameters.

A. Distribution of the Traffic in the Area

Analyzing the traffic density over the entire area of interest, we find that the amplitude X of the traffic density (meaning the values associated to the pixels) approximately follows a log-normal distribution (similar to the finding in [8]) with p.d.f.

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{(\ln x-\mu)^2}{2\sigma^2}}, \qquad (1)$$

and parameters $\mu = 0.68$ and $\sigma = 1.39$. The mean and variance of the traffic are $\mathbb{E}(X) = e^{\mu+\sigma^2/2} = 5.17$ Mbps/km^2 and $\mathrm{Var}(X) = (e^{\sigma^2}-1)e^{2\mu+\sigma^2} = 156$ Mbps/km^2, respectively. The fitted curve in Fig. 3 overestimates the data for lower traffic densities but is representative of the measurement, because the measurement underestimates the actual traffic (due to limited sampling). We use this result later on to investigate the relationship between the share of hot spot (or hot zone) traffic and the share of hot spot (or hot zone) area.

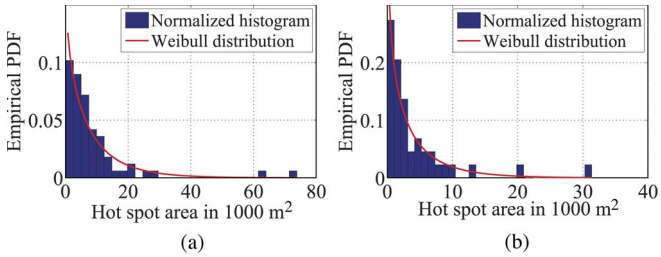

Fig. 4. Normalized histograms of the hot spot areas with Weibull fit for a threshold of (a) 5 times and (b) 10 times the mean traffic.

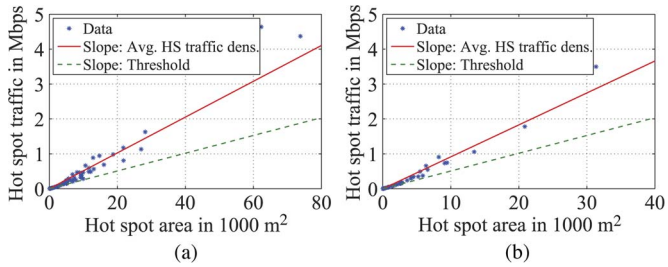

Fig. 5. Hot spot areas versus hot spot traffic volume for a threshold of (a) 5 times and (b) 10 times the mean traffic.

B. Hot Spot Densities

During investigations, it becomes obvious that the hot spot densities (and their respective sizes) are highly dependent on the threshold chosen. We observe that, for the given data, single isolated hot spots, which can be covered by small cells with coverage radii up to 150 m, can be identified with thresholds that are greater than five times the mean traffic density. In the remainder of the paper, we use two different thresholds, namely *five times and ten times the mean traffic*, to illustrate some important findings. For these thresholds, the average hot spot densities are 4 and 2.5 per km^2, respectively.

C. Distribution of Hot Spot Sizes

Fig. 4 depicts the histograms for the hot spot areas found for the two thresholds considered. We find that the probability distribution of the area sizes A of the hot spots (in units of 1000 m^2) can be approximated by a Weibull with p.d.f.

$$f_A(a;\lambda,k) = \frac{k}{\lambda}\left(\frac{a}{\lambda}\right)^{(k-1)} e^{(a/\lambda)^k}, \qquad (2)$$

and parameters $\lambda = 8.14$ and $k = 0.89$ for a threshold of five times the mean traffic, and $\lambda = 3.7$ and $k = 0.81$ for a threshold of ten times the mean traffic. In these cases, the mean hot spot area is 8600 m^2 (for 68 hot spots in total) and 4200 m^2 (for 42 hot spots in total), respectively.

D. Correlation between Hot Spot Sizes and Traffic Volumes

Fig. 5 shows the correlation between the size of the hot spots and the overall traffic volume in them. The (approximately) linear behavior observed indicates that there is a relatively high correlation between these quantities (see red line in Fig. 5).

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This document is a preprint of: H. Klessig, V. Suryaprakash, O. Blume, A. Fehske and G. Fettweis, "A Framework Enabling Spatial Analysis of Mobile Traffic Hot Spots," in IEEE Wireless Communications Letters (WCL), Oct 2014. DOI:10.1109/LWC.2014.2349520

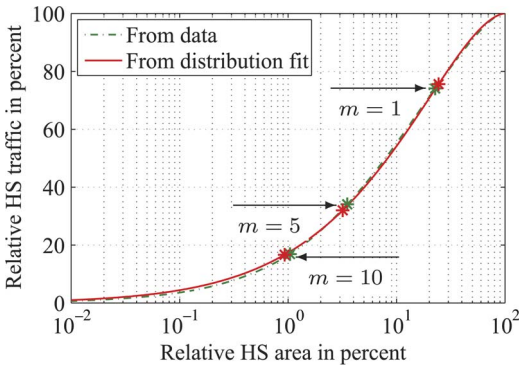

Fig. 6. Share of hot spot area versus share of hot spot traffic. Various lower thresholds t_1 (as m times the mean traffic) are inserted as asterisks.

Consequently, the mean traffic density in hot spot areas (slope of the red line) can be assumed to be independent of the hot spot size. The spread (or variance) of the points is determined by the choice of thresholds, which is basically the minimum slope of a line between the origin, (0, 0), and any sample point (see green dashed line). Fig. 5 also shows that smaller hot spot areas are more likely than larger hot spot areas (corresponding to the Weibull distribution). Note that the absolute values of the traffic may be scaled by considering an increase in mobile data traffic without any changes to the nature of the relationship illustrated, as pointed out in Section II-C.

E. Correlation Between the Share of Hot Spot Traffic and Area

Based on the result that the traffic is log-normally distributed with p.d.f. $f_X(x;\mu,\sigma)$ (1) and given two thresholds t_1 and t_2 $(t_2 > t_1)$, we can compute the fraction of the area with a traffic density between t_1 and t_2, and the fraction of total traffic covered as

$$P(t_1,t_2) = \int_{x=t_1}^{t_2} f_X(x)\,\mathrm{d}x \text{ and } T(t_1,t_2) = \frac{\int_{x=t_1}^{t_2} x f_X(x)\,\mathrm{d}x}{\int_{x=0}^{\infty} x f_X(x)\,\mathrm{d}x}, \quad (3)$$

respectively. The mean traffic density at these locations would be given by

$$T_{\mathrm{mean}}(t_1,t_2) = \frac{\int_{x=t_1}^{t_2} x f_X(x;\mu,\sigma)\,\mathrm{d}x}{\int_{x=t_1}^{t_2} f_X(x;\mu,\sigma)\,\mathrm{d}x}. \quad (4)$$

The slopes of the red lines in Fig. 5 correspond to $T_{\mathrm{mean}}(t_1,\infty)$ with t_1 being a threshold of five and ten times the mean traffic, respectively. For $t_2 \to \infty$, Eq. (3) provides the relationship between the share of hot spot area and the share of hot spot traffic (given a lower threshold t_1), which is depicted in Fig. 6. It shows that the heuristic curve (denoted as *From data*) matches the curve generated using the equations above. The information that can be gleaned from this figure can be explained as follows. A threshold with a value equal to the mean results in hot zones that cover 20% of the area, while containing 72% of the total traffic of the network. This, in fact, is a heuristic corroboration that the traffic in an area abides by the Pareto principle, which states that 80% of the effects come from 20% of the causes. The figure also reveals that single isolated hot spots, which are identified using a threshold of five times the mean traffic, cover about 3.2% of the total area and about 34% of the overall traffic volume (with an the average traffic density of $T_{\mathrm{mean}}(t_1,t_2) = 51.3$ Mbps/km^2).

IV. Conclusion

The tool and processing chain developed in this paper can be used for automatic hot spot detection and tracking. The use of this tool provides valuable insights into hot spot statistics for various cities and regions with respect to the number and sizes of traffic hot spots, as well as into efficient small cell deployment strategies. Since the results provided by this hot spot analysis framework may be subject to environmental changes, such as spatial traffic fluctuations during the course of a day or random noise, the pre-selection of meaningful call trace data and time intervals in which it is collected is essential. However, this is strongly determined by the user's intentions and the quality of call trace data itself. Based on the statistical findings with respect to the data given in this paper, small cells can be used to mitigate the expected traffic growth which exceeds today's macro cell capacity. With 2.5 small cells per km^2, 1% of the area can be covered and 17% of the traffic is offloaded. Even for offloading just twice the traffic, 3.2% of the area needs to be covered. Another doubling to 72% would require small cell deployment in 20% of the dense urban area, i.e., deploying more small cells in hot zones becomes increasingly costly in terms of the ratio of area covered and traffic demand serviced. In further work, we will apply this method to study the network energy efficiency gains of heterogeneous networks as a function of the area ratio covered by small cells.

Acknowledgment

The authors thank Mr. Hagen Freytag.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Taffic Forecast Update 2012–2017, San Jose, CA, USA, 2012, Tech. Rep.
[2] R. Razavi and H. Claussen, "Urban small cell deployments: Impact on the network energy consumption," in *Proc. IEEE WCNCW*, Apr. 2012, pp. 47–52.
[3] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.
[4] A. Kangas, I. Siomina, and T. Wigren, *Handbook of Position Location: Theory, Practice, and Advances*. Hoboken, NJ, USA: Wiley, 2011, ch. Positioning in LTE.
[5] E. Nan, X. Chu, W. Guo, and J. Zhang, "User data traffic analysis for 3g cellular networks," in *Proc. 8th Int. ICST Conf. CHINACOM*, Aug. 2013, pp. 468–472.
[6] G. Auer *et al.*, EARTH Project D2.3—Energy efficiency analysis of the reference systems, areas of improvements and target breakdown, 2011.
[7] T. Louail *et al.*, "From mobile phone data to the spatial structure of cities," *Sci. Rep.*, vol. 4, no. 5276, pp. 1–12, 2014.
[8] D. Lee, S. Zhou, and Z. Niu, "Spatial modeling of scalable spatially-correlated log-normal distributed traffic inhomogeneity and energy-efficient network planning," in *Proc. IEEE WCNC*, Apr. 2013, pp. 1285–1290.
[9] D. Lee *et al.*, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.
[10] L. Lucy, "An iterative technique for the rectification of observed distributions," *Astronom. J.*, vol. 79, pp. 745–754, Jun. 1974.
[11] W. H. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Amer.*, vol. 62, no. 1, pp. 55–59, Jan. 1972.
[12] E. V. Shikin and A. I. Plis, *Handbook on Splines for the User*, vol. 1. Boca Raton, FL, USA: CRC, 1995.
[13] S. L. Horowitz and T. Pavlidis, "Picture segmentation by a tree traversal algorithm," *J. ACM*, vol. 23, no. 2, pp. 368–388, Apr. 1976.

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.