Evaluating Urban Network Activity Hotspots through Granular Cluster Analysis of Spatio-Temporal Data

Jane Frances Pajo
Next Generation Technology,
Telenor Research
Fornebu, Norway
jane-frances.pajo@telenor.com

George Kousiouris
Dept. of Informatics and Telematics,
Harokopio University of Athens
Athens, Greece
gkousiou@hua.gr

Dimosthenis Kyriazis
Dept. of Digital Systems,
University of Piraeus
Piraeus, Greece
dimos@unipi.gr

Roberto Bruschi
[1] *DITEN – University of Genoa*
[2] *CNIT S2N National Laboratory*
Genoa, Italy
roberto.bruschi@unige.it

Franco Davoli
[1] *DITEN – University of Genoa*
[2] *CNIT S2N National Laboratory*
Genoa, Italy
franco.davoli@unige.it

Abstract**—Multi-access Edge Computing (MEC) is expected to play an essential role in enabling 5G (and beyond) technologies and services. This has driven numerous *micro-datacenter (μDC)* deployment studies in the literature, with a common goal of addressing the optimal μDC placement and dimensioning problems. Along this line, this paper aims at clustering subareas with similar network activity dynamics, to find a good hotspots' representation over the urban area. Leveraging common Machine Learning (ML) and statistics principles, the main contribution of this paper is two-fold: (i) the definition and selection of *dynamicity* features based on real telecommunications datasets; and (ii) the granular cluster evaluation and analysis based on *agglomerative hierarchical clustering*. Three feature sets (containing 20, 12 and 8 features, respectively) are evaluated at varying precision levels, showing interesting trends on the number of clusters, heatmaps and intra-cluster correlation. These could potentially provide some valuable indications on the placement and dimensioning of the μDCs.**

Keywords—Feature selection, Hierarchical clustering, MEC deployment, Network activity hotspots

I. Introduction

The 5G and beyond technologies and services have been recently pushing for the wide adoption of the Multi-access Edge Computing (MEC) paradigm [1], in order to accommodate the stringent requirements (such as low latencies, high connection density and seamless mobility support, among others [2]) of next-generation *verticals*. MEC deploys *micro-datacenters (μDCs)* towards the network edge to provide Cloud-like services much closer to end-users and their devices. Equipped with certain computing, networking and storage resources, the μDCs will be able to host both vertical application components and virtualized network functions (VNFs) that will be part of end-to-end services. Nevertheless, open issues on the optimal μDC placement and dimensioning need to be tackled prior to advancing towards large-scale deployments.

Towards this end, numerous works in the literature (e.g., [3]–[6], among others) have looked into different aspects of μDC deployment, such as the number of μDCs, as well as their locations and/or dimensioning, based on user demands and various quality and/or cost constraints. For instance, a comprehensive set of parameters (both quality of service (QoS)-related and not) is proposed in [3] for selecting μDC locations; while authors in [4] consider user location statistics to identify μDC potential locations, and analyze the impact of the number and dimensioning of the μDCs to the QoS. In [5], the authors proposed a mathematical model for finding the number and locations of 5G base stations and μDCs, by exploiting population statistics and considering the services' minimum base station distance constraints. Then, with the growth of Machine Learning (ML) applications in networking problems, the authors in [6] apply *k-means clustering* on base station coordinates to subdivide an urban area, and optimally place μDCs in each subarea through a *facility location* problem.

To the best of the authors' knowledge, there are still currently no μDC deployment studies in the literature evaluating urban network activity hotspots. In this respect, this paper takes the initial step with a granular cluster analysis of the spatio-temporal distribution of mobile network interactions over urban Milan. The goal is to cluster subareas that have similar dynamics and find a good network activity hotspots' representation over the urban area, which could potentially provide some valuable indications on the placement and dimensioning of the μDCs.

By leveraging common ML and statistics principles, the main contribution of this paper is two-fold: (i) the definition and selection of *dynamicity* features based on real telecommunications datasets; and (ii) the granular cluster evaluation and analysis based on *agglomerative hierarchical clustering*. The dynamicity features are derived from the subareas' network activity time series, and their *stationarized* by-products.

The remainder of this paper is organized as follows. Section II provides an overview of the dataset, and Section III describes the feature selection procedure adopted. The granular cluster evaluation and analysis are presented in Section IV, and finally, conclusions are drawn in Section V.

II. Dataset

This section provides a brief background on the dataset and the dynamicity features considered in this work.

This work was partially supported by the H2020 SPIDER, 5G-INDUCE and PolicyCLOUD projects, funded by the European Commission under Grants 833685, 101016941 and 870675, respectively.

A. The Milano Grid

Starting with the datasets from Telecom Italia's *Big Data Challenge*, we look into the network activity over Milan's urban area [7][8]. In particular, the Milano Grid [7] is composed of 10,000 square areas (i.e., a 100 x 100 grid), each one corresponding to an area of 235 m x 235 m. Fig. 1 provides a reference view on the Milano Grid coverage, also indicating the busiest square (near the *Duomo*) in the grid.

The telecommunications datasets, dated 2013, include a 2-month worth of mobile network data based on Call Detail Records (CDRs), which provides a set of time series related to users' SMS, calls and Internet activities, with values proportional to the corresponding type of network interaction instead of the actual load itself. Nonetheless, they can be exploited to analyze the dynamics of user-network interactions across the Milan urban area, such as network activity hotspots and unique/similar dynamic behaviors among the squares in the grid. More details on the dataset can be found in [9].

Supposing that the different types of network activities (i.e., *SMS-in*, *SMS-out*, *Call-in*, *Call-out* and *Internet traffic*) have the same weight, their values are summed up to generate a time series that would characterize each square. Hourly averages are further considered in this work, where the six samples in each hour of the original 10-minute interval time series are averaged.

It is important to note that while scaling the values in the dataset may be ideal to better represent the growth in network activity between 2013 and 2020, such intermediate preprocessing would not be necessary for the goal of this work.

B. Dynamicity Features

While each square area can have a unique behavior, it is intuitive to suppose that their geographical location and/or proximity with other areas can result in some similar dynamics. Hence, we explore different dynamicity features based on the areas´ network activity time series to cluster them accordingly.

Standard descriptive statistics can be extracted from time series data. Although they ignore the temporal succession among the samples, they still provide a general view on the dynamicity within the series. Five statistics are considered in this work, namely, the *minimum*, *maximum*, *mean*, *standard deviation* and *interquartile range* among the series samples.

Given the time series $\{x_k, k \in \mathbb{Z}\}$, the following stationarized series are also evaluated in an attempt to effectively capture dynamicity.

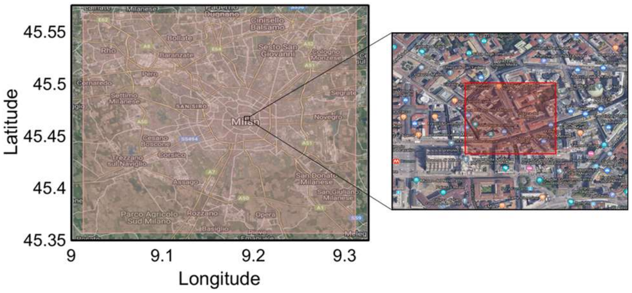


Fig. 1. The Milano Grid area and its busiest square area.

1) First order differences, $\{\dot{x}_k, k \in \mathbb{Z}\}$: The first derivatives along the time series' curve can be numerically approximated based on first order differences. With a step size equal to 1, the value at the k-th interval is given by

$$\dot{x}_k = x_k - x_{k-1} \tag{1}$$

2) Second order differences, $\{\ddot{x}_k, k \in \mathbb{Z}\}$: Similarly, its second derivatives can be numerically approximated from second order differences, and the value at the k-th interval is given by

$$\ddot{x}_k = \dot{x}_k - \dot{x}_{k-1} \tag{2}$$

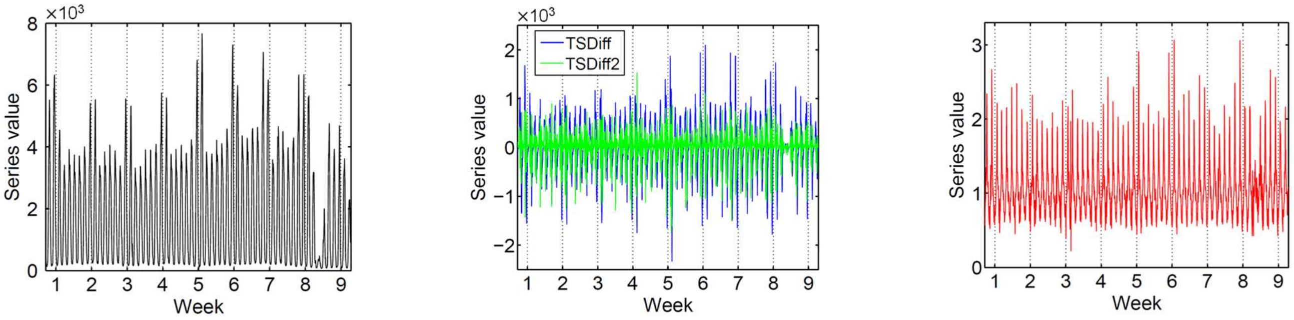
3) Link relatives, $\{\bar{x}_k, k \in \mathbb{Z}\}$: On the other hand, link relatives express the change between adjacent time series samples as ratios. The value at the k-th interval is given by

$$\bar{x}_k = \frac{x_k}{x_{k-1}} \tag{3}$$

Considering the busiest square area in the dataset, Fig. 2 illustrates a comparison between its original time series and the three stationarized series.

The corresponding descriptive statistics of the stationarized series are then extracted and considered as additional dynamicity features, besides those of the original time series. This results in a 20-tuple feature set {TSmin, TSmax, TSmean, TSstd, TSiqr, TSDiffmin, TSDiffmax, TSDiffmean, TSDiffstd, TSDiffiqr, TSDiff2min, TSDiff2max, TSDiff2mean, TSDiff2std, TSDiff2iqr, TSLinkRelmin, TSLinkRelmax, TSLinkRelmean, TSLinkRelstd, TSLinkReliqr} associated to each square area.

In order to evaluate the urban network activity hotspots, we seek to cluster square areas with similar behaviors based on their dynamicity features (or a subset thereof). Moreover, since the number of clusters is not known initially, agglomerative



(a) Original time series (b) First (TSDiff) and second (TSDiff2) order differences (c) Link relatives

Fig. 2. Original and stationarized series example for the network activity of the busiest square area in Telecom Italia's Big Data Challenge dataset [8].

hierarchical clustering is considered to obtain a cluster tree that can be cut at various precision levels to obtain a family of clusters at various granularities. Note that, depending on the chosen granularity, some clusters may contain a single coverage area, more specifically at high precision levels.

III. Dynamicity Feature Selection

Considering that the 20 features are all derived from the original time series and its stationarized by-products, some of them may be highly correlated and could lead to misleading results. Hence, it is only logical to deal with this potential problem through feature selection.

The coefficient of *determination* (R^2) is used to measure the similarities between dynamicity features to select a subset of more relevant features. Then, backward elimination is performed recursively until a stopping criterion (e.g., a maximum acceptable R^2 value) is reached. This procedure is detailed in ALGORITHM 1.

To summarize, let $Features$ be the initial set of dynamicity features, and $CoeffDet$ be the matrix of R^2 values, $R^2(i,j)$, corresponding to the feature pairs, (f_i, f_j), $i \neq j$. We define the R^2 threshold,

$$R_{th}^2 = floor(max[CoeffDet], 0.1) \qquad (4)$$

whose value is updated in each iteration. R_{th}^2 assumes a tenths value within $\{0.9, \dots, th_{min}\}$ as indicated by the *significance parameter* 0.1 ; the minimum value th_{min} serves as the stopping criterion of the algorithm. Starting with the feature pair corresponding to the maximum R^2 value, $(f_i^*, f_j^*)_{max}$, one of the two features is eliminated according to: (a) (N_i, N_j), the number of R^2 values greater than R_{th}^2 related to features f_i^* and f_j^*, respectively; and (b) (M_i, M_j), the maximum R^2 values, related to features f_i^* and f_j^*, respectively, when each is considered with the rest of the features f_k, $k \neq i^*, k \neq j^*, \forall k = 1, \dots, |Features|$. It is worth noting that the former criterion holds the higher priority, since removing the one having high R^2 values with respect to more features expedites the elimination process. This is repeated until all R^2 values in $CoeffDet$ are less than th_{min}, removing one feature from $Features$ in each iteration, along with its corresponding R^2 values in $CoeffDet$.

Setting $th_{min} = 0.6$, Fig. 3 shows how the maximum R^2 value in $CoeffDet$ and the number of features ($|Features|$) may vary at each iteration. It can be observed that the first 8 iterations assumed $R_{th}^2 = 0.9$, the following 4 iterations assumed $R_{th}^2 = 0.8$, then $\max[CoeffDet] < th_{min}$ is satisfied in the next and last iteration. Hence, the feature selection outcomes can be divided into three regions: $R^2 \geq 0.9$, $0.9 > R^2 \geq 0.8$, and $0.8 > R^2$, based on which we evaluate the initial 20-tuple feature set, together with the resulting 12- and 8-tuple feature sets in the granular cluster analysis. Particularly, the 12-tuple includes `Min`, `IQR`, `Mean_diff`, `Min_diff2`, `Mean_diff2`, `Min_linkrel`, `Max_linkrel`, `IQR_linkrel`, `Mean_linkrel`, `Max_diff`, `StdDev_diff2` and `Min_diff`; starting from this, the 8-tuple results in the elimination of the last four features.

ALGORITHM 1
Backward Elimination Process

```
$Features$
$CoeffDet \leftarrow \{R^2(i,j): i \neq j,\ i,j = 1, \dots, |Features|\}$
$th_{min}$
while max $[CoeffDet] \geq th_{min}$:
    $R_{th}^2 \leftarrow floor(\max[CoeffDet], 0.1)$
    $(f_i^*, f_j^*)_{max} \leftarrow (f_i, f_j)$ corresponding to max $[CoeffDet]$
    $N_i \leftarrow$ count $R^2(i^*, k) > R_{th}^2, k \neq i^*, k \neq j^*, \forall k = 1, \dots, |Features|$
    $N_j \leftarrow$ count $R^2(k, j^*) > R_{th}^2, k \neq i^*, k \neq j^*, \forall k = 1, \dots, |Features|$
    if $N_i > N_j$:
        Remove the $i^*$-th row and column from $CoeffDet$
        Remove $f_i^*$ from $Features$
    elseif $N_i < N_j$:
        Remove the $j^*$-th row and column from $CoeffDet$
        Remove $f_j^*$ from $Features$
    else:
        $M_i \leftarrow max[R^2(i^*, k)], k \neq i^*, k \neq j^*, \forall k = 1, \dots, |Features|$
        $M_j \leftarrow max[R^2(k, j^*)], k \neq i^*, k \neq j^*, \forall k = 1, \dots, |Features|$
        if $M_i > M_j$:
            Remove the $i^*$-th row and column from $CoeffDet$
            Remove $f_i^*$ from $Features$
        elseif $M_i < M_j$:
            Remove the $j^*$-th row and column from $CoeffDet$
            Remove $f_j^*$ from $Features$
        else:
            $k^* \leftarrow$ randomly choose between $i^*$ and $j^*$
            Remove the $k^*$-th row and column from $CoeffDet$
            Remove $f_k^*$ from $Features$
return $Features$
```

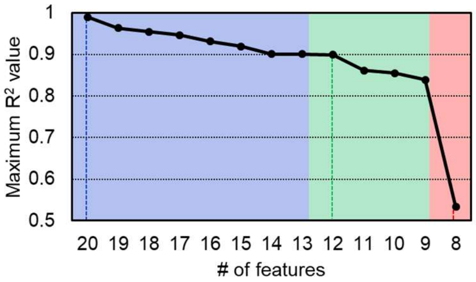


Fig. 3. Maximum $\boldsymbol{R^2}$ value in each iteration of the backward elimination process.

IV. Granular Cluster Evaluation and Analysis

An agglomerative hierarchical cluster tree is built for each of the aforementioned feature sets based on Matlab's implementation of the *Ward's minimum variance method* [10]. The trees are then evaluated in terms of the number of clusters, heatmaps and intra-cluster correlation, by cutting them at various precision levels.

A. Number of Clusters

A set of clusters can be obtained by cutting a cluster tree at the desired precision level. In this work, the cutoff points are given as percentages of the maximum Ward's distance in the tree. Evaluating cutoff distances at 1%, 2%, 5%, 10%, 20% and 50%, Fig. 4 illustrates how the number of clusters varies when considering the 20-, 12- and 8-tuple feature sets as input.

For instance, when cutting the cluster trees at 1% of the

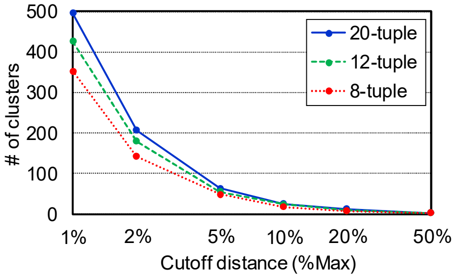


Fig. 4. Number of clusters when cutting the cluster trees (that result with the **20**-tuple, **12**-tuple and **8**-tuple feature sets) at various cutoff distances.

maximum Ward's distances, the 20-, 12- and 8-tuple feature sets result in 496, 426 and 353 clusters, respectively. More clusters naturally result with bigger feature sets since the relationships among data samples are captured with more detail. However, as previously anticipated, it is necessary to ensure that the feature sets do not include highly correlated features to avoid misleading results. As the precision level decreases (i.e., by admitting higher cutoff distances), the three cases tend to converge, resulting in 3 clusters at a cutoff distance of 50%.

Similarly, it is interesting to note that evaluating the three feature sets with the well-known k-means clustering method (e.g., either using the *Calinski-Harabasz* [11] or the *Davies-Bouldin* [12] indexes) for $k \in [2, \ldots, 500]$ yields around 2~3 clusters as optimal values of k. At such low precision levels, the clusters can be used to indicate, for instance, the urban core versus the suburbs (versus the agricultural areas). Nonetheless, a low intra-cluster correlation can be expected, since each cluster aggregates a substantially high number of square areas with highly heterogeneous network activity dynamics. Furthermore, simply tuning the k parameter of k-means does not provide a fine-grained control on the desired precision level as the agglomerative hierarchical clustering.

B. Heatmaps

Based on the average of the square areas' peak levels of network interaction within each cluster, Fig. 5 shows the resulting heatmaps over the Milano Grid area, when the 20-tuple, 12-tuple and 8-tuple cluster trees are cut at 1%, 5% and 20% of their corresponding maximum Ward's distances. Among a hundred possible heatmap color bins, around 1~3 cluster/s has/have been mapped with the same color in the figure.

Looking at Figs. 5a-c, it can be observed that cutting the trees at 1% of their maximum Ward's distances result in quite similar heatmaps, indicating a relatively dense hotspot within the urban core. The hotspots in the three cases consist of a single cluster, which in turn contains a single square area (i.e., the busiest one on the grid shown in Fig. 1). At this precision level, multiple other clusters also contain a single square area, accounting for around 14%, 9% and 7% of the total number of clusters for the 20-, 12- and 8-tuple feature sets, respectively. Intuitively, too much precision does not provide valuable indications for large-scale μDC deployments in this case. As the precision level is decreased, the number of hotspots and their coverage increases, as illustrated in Figs. 5d-i. At the same time, the percentage of single area clusters also decreases with the precision level, resulting in fewer, yet bigger clusters. In fact, a cut-off distance of 20% does not result in such single area clusters any longer.

It can also be observed how the hotspots resulting from the 20-tuple feature set are densely formed within/around the urban core. Then, as the number of features is decreased to 12 and 8, the number of hotspots gradually increases and spread across the Milano Grid area. Furthermore, increasing the cutoff distance for a given feature set results in the hotspots' wider coverage.

C. Intra-cluster Correlation

One way to validate the similarity between square areas in each cluster is through time series correlation. Particularly, considering the mean value of the coefficient of *correlation* (R) among pairs of square areas in each cluster, Fig. 6 shows the empirical *cumulative distribution function (CDF)* for the 20-, 12- and 8-tuple feature sets, when cutting the cluster trees at 1%, 5% and 20% of their corresponding maximum Ward's distances.

It is foreseen to achieve better intra-cluster correlation when utilizing more dynamicity features, as it can be observed when cutting the cluster trees at 1% and 20% of their maximum Ward's distances. However, it is interesting to note that at an intermediate cutoff distance of 5%, the 8-tuple feature set has resulted in better intra-cluster correlation than the 20- and 12-tuple feature sets, as indicated by the CDFs. Indeed, the heatmap in Fig. 5f seems to show the best mapping of hotspots with respect to Figs. 5d-e. Based on these results, it is evident that a joint optimization of both the feature set and the cut-off distance is necessary to obtain the best representation of urban network activity hotspots.

As a final remark, a heatmap such as the one in Fig. 5f can be exploited towards optimal μDC placement and dimensioning in large-scale deployments. For instance, the hotspots can give indications on the number and locations of μDCs, while the heatmap colors (corresponding to the average of the square areas' network interaction peaks within each cluster) can give indications on the dimensioning.

V. Conclusion

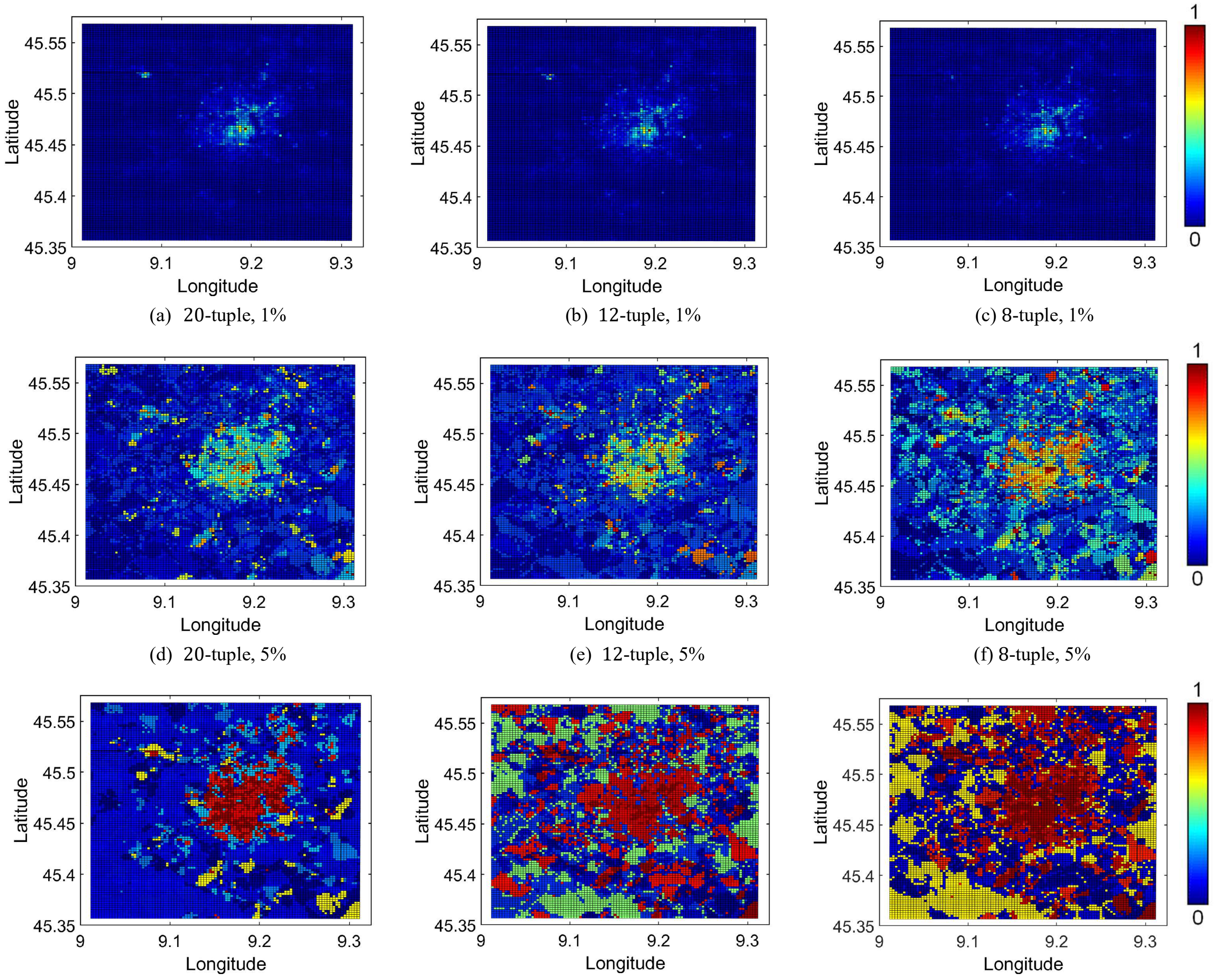
The MEC paradigm is expected to play an essential role in enabling 5G and beyond technologies and services. This has driven numerous μDC deployment studies in the literature. Along this line, this paper evaluated urban network activity hotspots through a granular cluster analysis of the spatio-temporal distribution of mobile network interactions over urban Milan, leveraging on common ML and statistics principles.

Three feature sets are evaluated with agglomerative hierarchical clustering, with the goal of clustering square areas that have similar dynamics. Their corresponding cluster trees can be cut at various granularities, allowing for a fine-grained control on the desired precision level.

Evaluation results show interesting trends on the number of clusters, urban network activity hotspots mapping and intra-cluster correlation, pointing out that: (i) too much precision does not necessarily give good results; and (ii) a joint optimization of both the feature set and the cut-off distance is necessary to obtain the best representation of urban network activity hotspots. These could potentially serve as valuable indications on the μDC placement and dimensioning in large-scale deployments.

References

[1] "Mobile Edge Computing (MEC); Framework and Reference Architecture," ETSI GS MEC 003 v1.1.1, Mar. 2016. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/01.01.01_60/gs_MEC003v010101p.pdf.



(a) 20-tuple, 1%

(b) 12-tuple, 1%

(c) 8-tuple, 1%

(d) 20-tuple, 5%

(e) 12-tuple, 5%

(f) 8-tuple, 5%

(g) 20-tuple, 20%

(h) 12-tuple, 20%

(i) 8-tuple, 20%

Fig. 5. Heatmaps over the Milano Grid, when cutting the **20**-tuple, **12**-tuple and **8**-tuple cluster trees at 1%, 5% and 20% of the maximum Ward's distances.

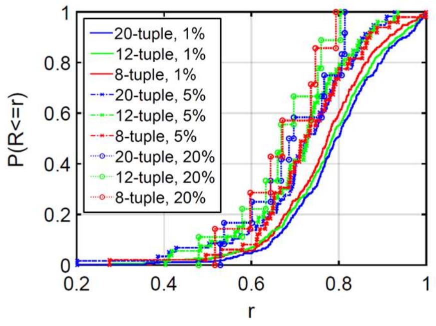


Fig. 6. Empirical CDF of the mean ***R*** value in each cluster, when cutting the cluster trees at 1%, 5% and 20% of the maximum Ward's distance.

[2] "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," ITU-R Rec. M.2083-0. Sept. 2015. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDFE.pdf.

[3] A. Santoyo-González and C. Cervelló-Pastor, "Edge Nodes Infrastructure Placement Parameters for 5G Networks," in Proc. 2018 IEEE Conf. Standards Commun. Netw. (CSCN), Paris, France, 2018, pp. 1-6.

[4] V. Burger et al., "Load Dynamics of a Multiplayer Online Battle Arena and Simulative Assessment of Edge Server Placements," in Proc. 7th Int. Conf. Multimedia Syst. (MMSys), Klagenfurt, Austria, May 2016, no. 17.

[5] J. Martín-Pérez et al., "Modeling Mobile Edge Computing Deployments for Low Latency Multimedia Services," in *IEEE Trans. Broadcasting*, vol. 65, no. 2, pp. 464-474, June 2019.

[6] U. Paul et al., "Traffic-profile and Machine Learning based Regional Data Center Design and Operation for 5G Network," in *J. Commun. Netw.*, vol. 21, no. 6, pp. 569-583, Dec. 2019.

[7] Telecom Italia, "Milano Grid," Harvard Dataverse, May 2015, [Online]. Available: https://doi.org/10.7910/DVN/QJWLFU.

[8] Telecom Italia, "Telecommunications - SMS, Call, Internet - MI," Harvard Dataverse, May 2015, [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV.

[9] G. Barlacchi et al., "A Multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino," in *Scientific Data*, vol. 2, Oct. 2015, Art. no. 150055.

[10] J. H. Ward, Jr., "Hierarchical Grouping to Optimize an Objective Function," in *J. Am. Stat. Assoc.*, vol. 58, no. 301, 1963, pp. 236–244.

[11] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," in *Commun. Stat.*, vol. 3, no. 1, 1974, pp. 1–27.

[12] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, Apr. 1979, pp. 224–227.