User mobility inference and clustering through LTE PDCCH data analysis

Pau Batlle Franch
UPC BarcelonaTech
Barcelona, Spain

Antoni Josep Eritja Olivella
UPC BarcelonaTech
Barcelona, Spain

Ramón María García Alarcia
UPC BarcelonaTech
Barcelona, Spain

Paolo Dini
CTTC/CERCA
Barcelona, Spain
paolo.dini@cttc.es

***Abstract*—The high penetration of mobile services provides an ample set of data generated by users and network elements. The analysis of such data yields insights on the behaviour of users and their experienced quality, and can be used by mobile operators to improve their mobile networks. In this paper, we design methods to infer user mobility patterns, estimate their channel quality and cluster them based on their Modulation and Coding Scheme (MCS) time evolution. In detail, we propose: i) a mapping between MCS and SNR, useful to assess the quality of the transmissions, ii) a method for deriving users' approximate velocities and categorising user mobility, and iii) a hierarchical clustering algorithm using Dynamic Time Warping, able to generate meaningful user clusters according to communication length and quality. We apply those proposed methods to real traces collected from more than one week of observations of three operative base stations in Spain. We observe that our solutions successfully provide relevant information about users' mobility and their channel quality, making them suitable for improvements in the understanding and planning of LTE resources by mobile network operators.**

***Index Terms*—Channel Quality Analysis, Channel Quality Time Evolution, Clustering, Data Analysis, Dynamic Time Warping, Pdcch, Time Series Analysis, Time Series Clustering, Users Mobility, Wireless Traffic Analysis, Wireless Traffic Characterization.**

I. Introduction

The use of mobile data communications has become massive in the era of smartphones and the Internet of Things. With the increasing pressure of more users and devices connected than ever before and a growing concern on reducing energy consumption to the highest possible extent, data analysis has become a cornerstone in the optimisation and engineering of future mobile equipment and networks [1]. Applying data analysis techniques to LTE networks enables proactive and self-organised network management. For example, it is important to be able to predict when a base station will be overloaded or underloaded to properly manage radio resources among users or trigger energy saving modes, respectively. For that purpose, it is key to study user mobility patterns and estimate the quality required by the requested services.

Among different data analysis techniques, Machine Learning has been extensively used for different mobile communication tasks (see [2] for a complete survey). In particular, unsupervised clustering techniques have been used as part of pipelines for user classification tasks [3] [4] [5], and also to learn mobility patterns [6]. In [7], an algorithm to group base stations in different clusters is proposed using aggregated data, and typically K-means [8] and Self-Organising Maps (SOM) [9] have been used for these tasks.

In this work, we use the information on Modulation and Coding Scheme (MCS) included into the DCI messages of the Physical Downlink Control Channel (PDCCH) from a LTE operative network in Spain to infer user mobility patterns, estimate their channel quality and cluster them based on their MCS time evolution. The main contributions of this paper are as follows. First, we propose a mapping between MCS and SNR and find a relation between the number of transmitting users in a base station and the quality they experience. Second, we use the approximate SNR values to infer users' mobility patterns. Finally, a distance based hierarchical clustering framework is used to identify users receiving similar communication patterns over time. The proposed method has the main advantage of being able to cluster time series of different length and uses a distance metric suited for time series specifically. Furthermore, the adoption of hierarchical clustering [10] implies that the number of clusters can be effectively chosen *after* the clustering procedure is done.

This article is structured as follows. In Section II, we present the data used along with some considerations regarding data preprocessing. In Section III, we present the theory and procedures used for mobility inference and user clustering, and the results of applying such procedures to the presented data appear in Section IV. Finally, a conclusion is given in Section V.

II. Data and preprocessing considerations

The dataset used in this paper was part of the EU-MSCA-SCAVENGE data challenge [11]. The data consists of traces recorded from the PDCCH (Physical Downlink Control Channel) of three eNBs (eNodeBs), henceforth referred as stations "A", "B", and "C", located in three different areas of the city of Barcelona, with over a week of observations for each station. The PDCCH is the channel that a LTE eNB uses, among other purposes, to handshake the resource scheduling information to transmitting UEs (User Equipments) both in uplink and downlink.

The dataset is composed of 4 columns: Timestamp, Transmission direction (UL/DL), Modulation and Coding Scheme

(MCS) index (from 0 to 31) and Temporal User Identifier (C-RNTI).

The MCS is a number (in LTE, ranging from 0 to 31) that is related to a pair of Modulation and Code Rate. LTE uses different digital modulations (QPSK, 16-QAM, 64-QAM, 256-QAM), and different code rates to adjust channel quality between an eNB and a user at any time, so that re-transmissions of data are minimised [12]. MCS indices are allocated by the eNB as a function of the Channel Quality Indicator (CQI) sent by the UE. The CQI depends on the SNR experienced by the user, which, among other factors, is generally correlated with the distance between the eNodeB and the UE. Hence, the information on the assigned MCS, may be used to infer user mobility patterns [12].

The C-RNTI is a LTE 16-bit number which, among its several purposes, helps in identifying a UE in shared channels such as the PDCCH [12]. Being temporary, C-RNTIs do not uniquely identify users: since there is only a limited number of RNTIs, two users who communicate sufficiently distant in time might share the same C-RNTI value. The expiration time of C-RNTIs is an eNB vendor specific parameter and, therefore, unknown in the dataset. The analysis of the distribution of typical user communication lengths for different values of this threshold suggested the reasonable value of 1 hour. This avoids the unrealistic cases of too many very short or very long communications. After the data preparation procedures, we can associate each user i with a vector $\overrightarrow{\text{MCS}}_i = (\text{MCS}_i^1, ..., \text{MCS}_i^{l(i)})$, being MCS_i^k, the k^{th} MCS measurement of the user i and $l(i)$ the number of measurements in the users' session. Each MCS vector comes with an auxiliary vector of timestamps $(t_i^1, ..., t_i^{l(i)})$. Two users i and j might not have the same communication lengths, so in general $l(i) \neq l(j)$ for $i \neq j$.

III. Procedures

A. SNR and mobility analysis

The approach to infer user mobility relies on the estimation of users' SNR, which, in turn, can be used to estimate their velocity by analysing its variation over time. We propose a mapping from MCS to SNR consisting of two steps: A MCS-CQI mapping based on LTE specifications [12] and a CQI-SNR mapping partly based on [13].

Table I shows the proposed MCS-SNR mapping. We propose a mapping based on the assumption that the SNR follows a normal distribution, and thus, that it is more reasonable to find, in each modulation order, medium CQI levels than low or high CQI levels. Hence, we propose a mapping which overrepresents medium CQI levels (which are mapped to more MCS levels) and underrepresents low and high CQI levels (which are mapped to less MCS levels), as it can be seen in Table I.

The mapping from CQI to SNR depends on the eNB configuration [13] (whether a single antenna, diversity, Multiple-input Multiple-output -MIMO-, etc. are used). Thus, two SNR values are provided for each CQI: "SNR_{best}" which is the SNR in the best configuration scenario, "$\text{SNR}_{\text{worst}}$" which is the SNR in the worst configuration scenario, as the Table depicts.

MCS	CQI	SNR_{best} (dB)	$\text{SNR}_{\text{worst}}$ (dB)
0	1	2.00	-7.00
1-2	2	4.05	-5.00
3-5	3	6.00	-3.15
6-7	4	8.00	-1.00
8-9	5	10.00	1.00
10	6	11.95	3.00
11-13	7	14.05	5.00
14-17	8	16.10	6.90
18-20	9	17.90	8.90
21	10	20.50	10.85
22	11	22.45	12.60
23-24	12	23.45	14.35
25-26	13	25.00	16.15
27	14	27.30	18.15
28	15	29.60	20.00
29	3	6.00	-3.15
30	8	16.10	6.9
31	12	23.45	14.35

TABLE I: Proposed mapping between MCS and SNR using CQI. Since the mapping between MCS and CQI is not unique, a design choice to overrepresent medium CQI values in opposed to extremes was made. The SNR depends on unknown parameters, hence, a confidence interval between $\text{SNR}_{\text{worst}}$ and SNR_{best} is given.

The velocity of user i at its measurement j, v_i^j, is computed using Equation (1). We first compute users' SNR vectors $\overrightarrow{\text{SNR}}_i$ from the users' MCS vectors $\overrightarrow{\text{MCS}}_i$, by applying Table I. The variation of SNR (ΔSNR_i^j) and of time ($\Delta t_i^j = t_i^j - t_i^{j-1}$) are needed. To guarantee that the SNR variation value is significant (i.e., that the value accounts for the SNR margin of uncertainty), it is computed as $\Delta\text{SNR}_i^j = \max(0, \text{SNRworst}_{i+1}^j - \text{SNRbest}_i^j, \text{SNRworst}_i^j - \text{SNRbest}_{i+1}^j)$. The ratio between these two variables is then multiplied by the constant d_{1dB} to obtain the user's velocity, being d_{1dB} the variation of the distance for a change of 1 dB in the SNR. d_{1dB} is derived using Equation (2) for a given frequency. In this case, $f = 2600\,\text{MHz}$.

$$v_i^j = \frac{d_{1dB}}{\Delta t_i^j}\Delta\text{SNR}_i^j = \frac{0.010302m}{\Delta t_i^j}\Delta\text{SNR}_i^j \quad (1)$$

$$\text{FSL} = \left(\frac{4\pi d}{\lambda}\right)^2 = 1dB \Rightarrow d = \frac{\sqrt{FSL}\lambda}{4\pi} = \frac{\sqrt{FSL}c}{4\pi f} \quad (2)$$

where FSL is the Free-Space Loss attenuation [dB], c is the speed of light [m/s], f is the signal frequency [Hz] and d is the distance change for a given attenuation variation which becomes d_{1dB} when FSL is set to 1 dB [m].

B. User clustering based on MCS patterns

The objective of our clustering procedure is to separate the different users' $\overrightarrow{\text{MCS}}_i$ vectors into groups based on their communication behaviour (i.e., session length and MCS dynamics). It is important to recall that our dataset represents the different MCS values of a given C-RNTI during the duration

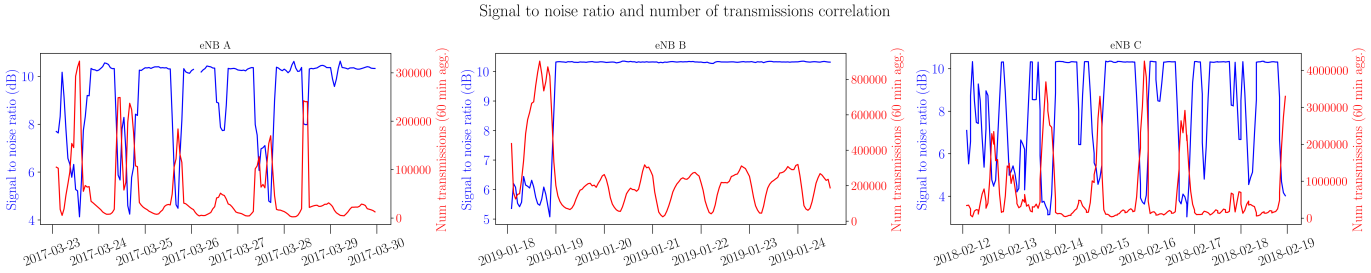


Fig. 1: Time evolution of average SNRs (in blue) for all users and number of communications (in red) at each time stamp for the three eNBs. On average, a higher number of communications leads to lower SNRs.

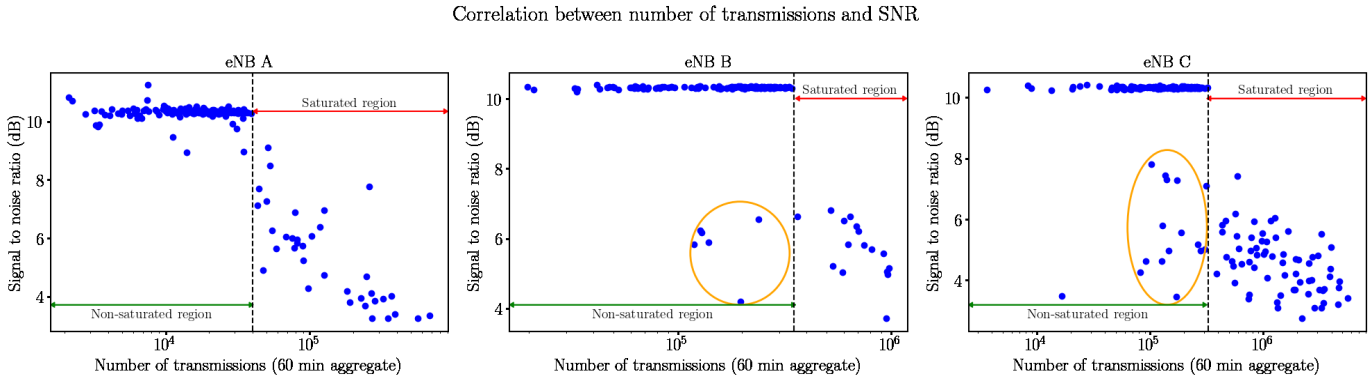


Fig. 2: Correlation between SNR and log of number of communications for each of the eNBs. One point corresponds to 60 min. Users mostly experience high SNR while the number of communications is low enough, and the SNR starts to decrease on average after a threshold. Quality was also observed to decrease due to other external factors (orange circles).

of the user's communication. Hence, the several time-series obtained differ in length and it is impossible to apply the most classical clustering algorithms (K-means, SOM) without modifying the vectors first. Instead, here we employ a pair-wise distance based clustering using *Dynamic Time Warping* (DTW), as the distance function. DTW is able to calculate distances between sequences of different length, and has been widely used before in the time series analysis literature [14] [15] [16]. The pairwise distance matrix obtained from applying DTW to the observations can then be fed into a hierarchical clustering algorithm, which defines a hierarchy of clusters, so the number of clusters can be defined *a posteriori*. A linkage matrix is created to store the clusters at each step. Here, Ward's variance minimization algorithm [10] is used to calculate the minimum distance between the observations and the clusters. After that, the number of clusters is decided and the cluster assigned to each observation is obtained. The whole procedure is presented in Algorithm 1. In the pseudocode, linkage refers to the hierarchical clustering algorithm linkage function implemented in libraries for popular programming languages such as *Python*, *Matlab* or *Mathematica*, in which the usage of Ward's algorithm needs to be specified. For any $1 \leq k \leq n =$ Number of users, k clusters can be retrieved using the output linkage matrix Z.

Algorithm 1: Clustering procedure

Input : $\{\overrightarrow{\text{MCS}}_i\}$, set of MCS observations of different lengths for different users
Output: Hierarchy of clusters and dendrogram, which can be used to assign clusters to each observation

$n \leftarrow$ Number of users to cluster
$M \leftarrow$ Matrix(n,n)
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, i$ **do**
    $M_{ij} \leftarrow \text{DynamicTimeWarping}(\overrightarrow{\text{MCS}}_i, \overrightarrow{\text{MCS}}_j)$
Z = linkage(M)
return Z

IV. Results and Discussion

All of the data analysis and procedures were implemented in *Python*, mainly relying on the data manipulation library Pandas, and machine learning and scientific computing libraries Scipy and Scikit-learn.

A. SNR and mobility analysis

Figure 1 shows the average of users' SNR and the number of messages exchanged at every time interval in the dataset for each monitored eNB. This shows the impact of the number of

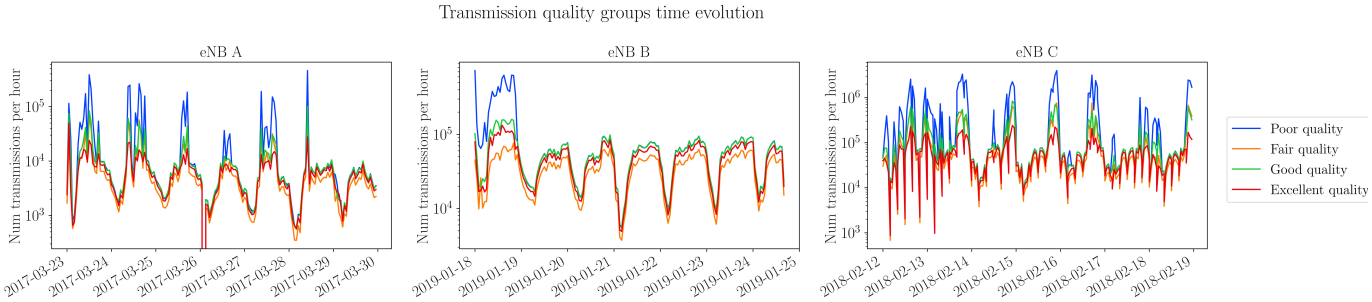


Fig. 3: Evolution of communication types by quality over time. We observe the saturation phenomena explained before: poor communication mainly arises when the number of transmissions is high, while communication quality is mostly high when the number of communications is low. Data is plotted in 60-min intervals.

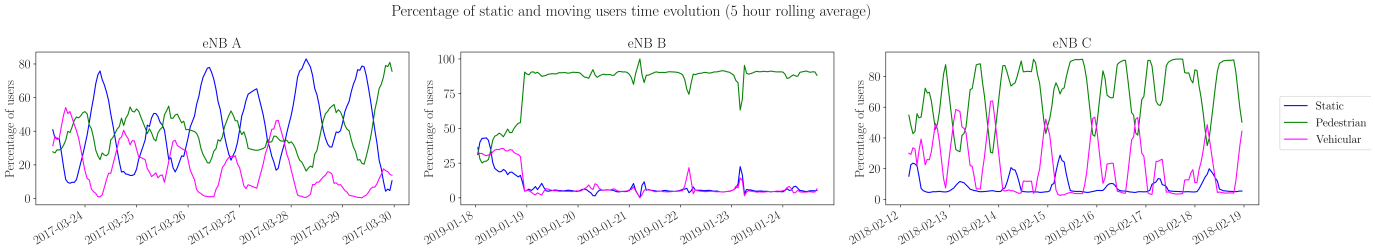


Fig. 4: Evolution of communication type percentages by mobility types over time, including static (blue), pedestrian (green) and vehicular (purple). The three eNB have different behaviour: while in eNB A (and to a lesser extent, in eNB C) the existence of pedestrian users depends a lot on the time of the day, for eNB B we observe that users are mostly pedestrian at all times.

competing users on their SNR. The correlation between the two phenomena is quite clear: a higher number of competing users leads to a worse SNR. Moreover, it is also possible to note a regular pattern during the days: a higher number of messages is always exchanged around midday/early afternoon in eNB A and eNB C, whereas the peaks are during the evening in eNB B.

In Figure 2, a different visualisation of the same SNR analysis is presented. In this case the average SNRs are plotted against the logarithm of the number of incoming communications in a 2D scatter plot in which each point represents 60 minutes of transmissions. Given the limited spectrum available in wireless communications systems, only a limited number of users can be served without degrading their service quality. Thus, there is a region in which the SNR does not primarily depend on the number of users, and a point from which the SNR starts to drop with an increasing number of users. This point defines the beginning of what we define as saturation region. In Figure 2, the left side is the area in which an increase in the number of messages does not lead to a decrease in the SNR of the user communications, which in general remains constant. On the contrary, on the right side, the SNR starts to decrease with $\log_{10}(n_t)$ where n_t is the number of transmissions. The limit is at around 4×10^4 transmissions for eNB A and around 3.5×10^5 for eNB B and eNB C. However, for eNB B and eNB C, some points are spotted inside the non-saturated region well beyond the mean SNR of the mentioned region (circled in orange). This means that even though no quality degradation is due to the amount of users in the eNB, as the number of transmissions is theoretically supported by the base station, there is something impairing the channel quality. eNB B and eNB C might be covering a more complicated area than eNB A in terms of morphology, where damaging conditions for wireless transmission are present (e.g. narrower streets, higher number of urban obstacles, etc.).

In order to understand better the evolution of the user transmission quality over time, we have introduced four quality categories: Poor, Fair, Good, and Excellent. A Poor quality transmission experience is defined to be the one in which SNR_{worst} is below 0 dB. For Fair quality, SNR_{worst} is at or above 0 dB and below 6 dB. Good quality represents an SNR_{worst} value equal to or above 6 dB and below 12 dB. Finally, an Excellent quality means that the SNR_{worst} meets or exceeds 12 dB. This can be translated to MCS using Table I. Each user MCS sample MCS_i^j is mapped into its corresponding quality tag defined in the above, and Figure 3 shows the evolution of quality tags over time. When the number of messages exchanged is high, the predominant quality is Poor: the interference between users becomes an important factor, and SNR is lower. When a lower number of users are competing, Good and Excellent quality are more predominant. This can be more clearly noticed for eNB B.

Finally, the average speed of the users is calculated based on the velocity estimation procedure explained in the previous section. With this value, a user is classified into static or

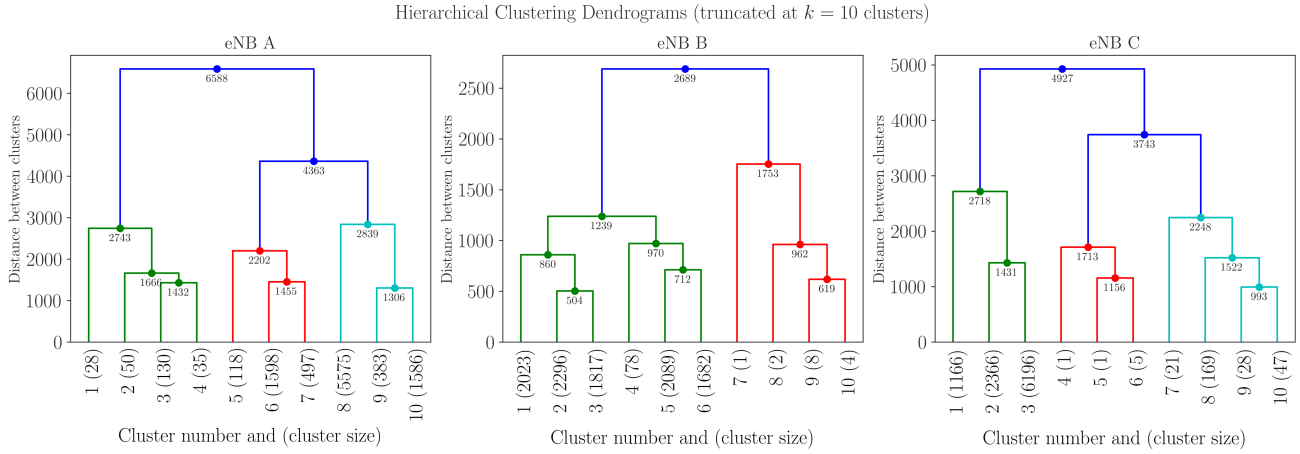


Fig. 5: Dendrogram of hierarchical clusters based on MCS patterns, truncated at $k = 10$ clusters for clarity. This figure represents the bifurcation order in the hierarchical clustering algorithm in order to obtain the final 10 clusters. The y axis represents how much distance betweeen clusters exists at a given bifurcation, and the leaves are each of the clusters, with the size in number of observations in between parenthesis.

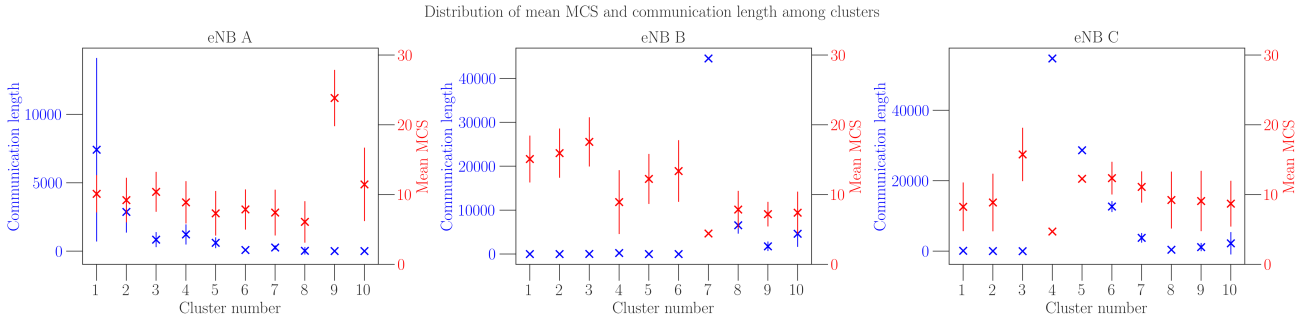


Fig. 6: Distribution of mean MCS and communication length among the $k = 10$ clusters created for each eNB using hierarchical clustering. Each point represents a cluster, with the cross signaling the average of MCS (red) and communication length (blue) and the vertical bars one standard deviation away from the average.

moving, depending if its velocity is close to 0 m/s or greater than 0 m/s, respectively. Inside the moving category, users are divided into 'pedestrian' and 'vehicular' depending on their estimated velocity being below or above 10 km/h (2.8 m/s). Table II depicts the distribution of users into the three mobility categories. The distributions of eNB B and eNB C are quite similar, with the majority of users being classified as pedestrian. In eNB A, where the amount of users is also quite smaller, most of them are classified as vehicular.

eNB ID	Num. of users	% static	% pedestrian	% vehicular
A	41 971	18.86	31.77	49.37
B	1 274 283	6.49	86.17	7.34
C	1 377 989	5.67	72.47	21.86

TABLE II: Percentage of static, pedestrian, and vehicular users for each eNB.

Figure 4 shows how the amount of users in each mobility category evolves over time. Again, differences are spotted between base stations. For instance, the mobility distribution is volatile in eNB A and eNB C, whereas it is quite consistent in time in eNB B.

Due to the simplicity of its underlying FSL model it must be noted that this mobility assessment procedure might provide inaccurate data in specific situations of high channel quality disruptions. However, it is simple to implement and the results presented above have shown that it is useful when performing general mobility assessments.

B. User clustering based on MCS patterns

The clustering procedure in Algorithm 1 is applied separately to random subsets of $N = 10000$ communications on each of the 3 eNB. We take $k = 10$ clusters for all cases to demonstrate results, but this number can be easily adjusted after the algorithm has run due to the notion of distance between clusters, making it useful for practical applications.

Figure 5 shows the resulting dendrograms (truncated after 10 clusters) for each case. The amount of users in each cluster and the distance between clusters are also plotted. The position on the y axis of every fork marks the distance between

clusters. If less than 10 clusters are needed, they can be merged starting from bottom to top. Colours represent proximity, with same-coloured clusters being closer between them than to other colours. In order to assess the validity of the clustering algorithm proposed, we plot the mean and standard deviation of the average communication quality and communication length for each cluster at each eNB in Figure 6. We observe that close clusters are similar in observation length, and the difference between them is the mean MCS. This means that the hierarchical clustering separates observations first based on length and then based on quality. This is in agreement with the fact that DTW assigns higher distances to vectors that have different length. This helps in separating outliers such as clusters 7-10 in eNB B and 4-6 in eNB C. Once these are separated, the clustering is able to distinguish between different MCS values for vectors of similar length, as seen, for example, in clusters 8-10 in eNB A, and also, 1-6 in eNB B and 1-3 in eNB C. This reasoning helps us verify that the performed clustering of users is meaningful and could be useful for additional downstream tasks such as classification, prediction, quick analysis of data from new eNBs, etc.

V. Conclusions

In this paper we have presented methods to analyse the mobility of users, evaluate their channel quality and cluster them into groups given a set of prerecorded PDCCH traces from operative LTE eNBs. The relationship between transmission quality and user mobility has been also studied. We have introduced a mapping between MCS and SNR to assess the quality of the transmissions, a method for deriving users' approximate velocities and categorising user mobility, and a hierarchical clustering algorithm using Dynamic Time Warping able to generate meaningful user clusters according to communication length and quality. The hierarchical clustering algorithm is additionally able to detect data outliers by separating them in very small clusters. Applying the proposed methods to data coming from three different real eNB we observed that they successfully provide relevant information and insights about users' mobility and their channel quality making them suitable for improving the understanding and planning the LTE resources by mobile network operators.

Acknowledgements

This work has received funding from Spanish MINECO grant TEC2017-88373-R (5G-REFINE), from Generalitat de Catalunya grant 2017 SGR 1195 and from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675891 (SCAVENGE). The authors wish also to thank Prof. Mario García-Lozano (UPC), for advanced insights about the behaviour of LTE, and Prof. Antoni Gelonch (UPC), for providing the authors access to a LTE test bench to help back assumptions regarding the dataset.

References

[1] R. Boutaba, Mohammad Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, and Oscar Caicedo Rendon. A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities. *Journal of Internet Services and Applications*, 9, 05 2018.

[2] Jessica Moysen and Lorenza Giupponi. From 4g to 5g: Self-organized network management meets machine learning. *Computer Communications*, 129:248–268, September 2018.

[3] Pietro Savazzi and Lorenzo Favalli. Dynamic cell sectorization using clustering algorithms. In *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*. IEEE, April 2007.

[4] Shutchon Premchaisawatt and Nararat Ruangchaijatupon. Enhancing indoor positioning based on partitioning cascade machine learning models. In *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, May 2014.

[5] Yunjuan Zang, Feixiang Ni, Zhiyong Feng, Shuguang Cui, and Zhi Ding. Wavelet transform processing for cellular traffic prediction in machine learning networks. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, July 2015.

[6] Juan Sánchez-González, Jordi Pérez-Romero, Ramon Agustí, and Oriol Sallent. On learning mobility patterns in cellular networks. In *IFIP Advances in Information and Communication Technology*, pages 686–696. Springer International Publishing, 2016.

[7] Longbiao Chen, Dingqi Yang, Daqing Zhang, Cheng Wang, Jonathan Li, and Thi-Mai-Trang Nguyen. Deep mobile traffic forecast and complementary base station clustering for c-RAN optimization. *Journal of Network and Computer Applications*, 121:59–69, November 2018.

[8] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[9] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[10] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

[11] Scavenge Group. Scavenge data. https://www.scavenge.eu/dissemination/data-challenge/. Accessed: 2019-02-01.

[12] ETSI. TS 36.213 Evolved Universal Terrestrial Radio Access (E-UTRA): Physical layer procedures, version 8.3.0 release 8. Technical specification, ETSI, 11 2008.

[13] Mohammad Kawser, Nafiz Imtiaz Bin Hamid, Md Nayeemul Hasan, M Shah Alam, and M Musfiqur Rahman. Downlink SNR to CQI mapping for different multiple antenna techniques in LTE. *International Journal of Information and Electronics Engineering*, 2:756–760, 09 2012.

[14] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978.

[15] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409, September 1981.

[16] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 2007.