Characterizing Mobile Service Demands at Indoor Cellular Networks

Stefanos Bakirtzis
Ranplan Wireless &
University of Cambridge

André Felipe Zanella
IMDEA Networks Institute &
Universidad Carlos III de Madrid

Stefania Rubrichi
Orange Innovation

Cezary Ziemlicki
Orange Innovation

Zbigniew Smoreda
Orange Innovation

Ian Wassell
University of Cambridge

Jie Zhang
Ranplan Wireless &
University of Sheffield

Marco Fiore
IMDEA Networks Institute

ABSTRACT

Indoor cellular networks (ICNs) are anticipated to become a principal component of 5G and beyond systems. ICNs aim at extending network coverage and enhancing users' quality of service and experience, consequently producing a substantial volume of traffic in the coming years. Despite the increasing importance that ICNs will have in cellular deployments, there is nowadays little understanding of the type of traffic demands that they serve. Our work contributes to closing that gap, by providing a first characterization of the usage of mobile services across more than 4, 500 cellular antennas deployed at over 1, 000 indoor locations in a whole country. Our analysis reveals that ICNs inherently manifest a limited set of mobile application utilization profiles, which are not present in conventional outdoor macro base stations (BSs). We interpret the indoor traffic profiles via explainable machine learning techniques, and show how they are correlated to the indoor environment. Our findings show how indoor cellular demands are strongly dependent on the nature of the deployment location, which allows anticipating the type of demands that indoor 5G networks will have to serve and paves the way for their efficient planning and dimensioning.

CCS CONCEPTS

• **Networks** → **Network performance evaluation**; *Network measurement*; *Network performance modeling*.

KEYWORDS

Indoor cellular networks, mobile service demands, explainable machine learning, traffic profiles

ACM Reference Format:
Stefanos Bakirtzis, André Felipe Zanella, Stefania Rubrichi, Cezary Ziemlicki, Zbigniew Smoreda, Ian Wassell, Jie Zhang, and Marco Fiore. 2023. Characterizing Mobile Service Demands at Indoor Cellular Networks. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3618257.3624807

IMC '23, October 24–26, 2023, Montreal, QC, Canada

1 INTRODUCTION

Understanding mobile network usage is an important task with manifold implications, in networking and beyond. The exploration, characterization, and modeling of mobile network traffic have a key role in developing and supporting an efficient wireless ecosystem [40]: understanding and forecasting traffic demands enables the proactive configuration of the wireless network [35] in order to accommodate the needs of end users and provide them with enhanced quality of service and experience [52, 57]. At the same time, the traffic generated by mobile services has become a primary source of insights on human activities, needs, and habits [20]; this has benefited research in diverse scientific disciplines, ranging from socioeconomic topics like unemployment [32], wealth distribution [43, 49, 9], social inequality and access to digital services and media [37, 50] to climate change [12, 10]. More recently, suitably processed mobile network traffic has been employed to track and prevent the spread of infectious diseases [25, 41] or to assess their impact on human habits [56].

Interestingly, all existing studies on characterizing and exploiting mobile traffic invariably focus on measurements collected in outdoor environments. This can be ascribed to the fact that cellular networks have been primarily intended as a technology for mobile, outdoor users. Yet, this is not true anymore, and main actors in the telco ecosystem forecast that about 80% of the future cellular data traffic will be generated in indoor environments [11]. As a result, fifth-generation (5G) and beyond (B5G) systems are anticipated to align with the emerging need to serve indoor users [1]: specifically, a number of major vendors and mobile network operators (MNOs) expect 5G/B5G systems to transition from a legacy "outside-in" coverage approach, where indoor coverage is provided by antennas located outdoor, to the deployment of native indoor cellular networks (ICNs) [11, 14, 27, 61]. ICN deployments will allow improving substantially the quality of service for indoor user equipments (UEs), and bring real competition to Wi-Fi technologies for the increasingly remunerative indoor market.

The emergence of pervasive ICNs makes it important to comprehend how such networks will be used. In this context, a considerable volume of research has focused on methods related to the modeling of radio propagation in indoor environments [46, 36, 4], the impact of the building layout on network performance [26, 59, 58, 6] or the efficient planning of ICNs [19, 55, 54, 5]. However, the

dynamics and distinguishing features of the traffic data generated by in-building radio access network components have not been studied yet. Unlike outdoor base stations (BSs) that tend to observe a general-purpose use, *i.e.*, serve concurrently numerous subscribers engaged with diverse activities during different daily endeavors, ICNs are expected to target more specific use cases. For instance, ICNs are deployed in underground subway and train stations to compensate for the limited coverage of the outdoor wireless network. Likewise, corporate offices are equipped with indoor BSs to provide enhanced and reliable communications to support the work of their employees [2], whilst their installation is necessary in exposition centers and stadiums in order to accommodate the concentrated high-traffic demands intertwined with social events. Therefore, the ICN traffic is reasonably influenced by the context in which it is generated, which highly depends on the indoor environment type and, by extension, on the kind of activities in which users are involved in it. Eventually, the traffic dynamics of ICNs are expected to differ significantly from those of legacy communication system outdoor BSs.

As mentioned though, and to the best of our knowledge, until now there has been no research to study thoroughly the characteristics of the traffic generated by ICNs. Unlike previous research in the field that focused on understanding and predicting the dynamics of the macro BSs traffic [52, 57, 18, 17, 28], our work dwells upon the intrinsic particularities of indoor cellular traffic. In light of the proliferation of indoor communication systems and the establishment of private networks, the results of our analysis provide novel insights that may support an improved design and operation of these networks, *e.g.*, via resource allocation, network slicing, caching, or energy adaptation schemes. In particular, our work sets forth the following contributions:

- Hinging on a countrywide ICN Internet measurement traffic data set, we define an appropriate transformation of the traffic data that enables probing the range of different Internet mobile service utilization profiles at indoor antennas. Then, employing an unsupervised learning approach, we designate that distinct service utilization clusters are inherent in indoor communication systems. This rich and diverse behavior of Internet services has not been unveiled before, and as we also demonstrate, does not align with that of outdoor legacy communication systems.
- To interpret the clustering results and delve into the essence of the different clusters, we leverage techniques from the field of explainable machine learning (ML). In particular, we render the unsupervised learning results interpretable by employing the Shapley additive explanations (SHAP) framework [33]. This enables the identification of the most important features for each cluster and consequently allows us to expound on the most important as well as on under-utilized Internet service types of each cluster.
- We expose that there is a strong connection between the clusters individuated by our analysis and the indoor environment type. In particular, we show that the same mobile applications manifest very heterogeneous behaviors between ICNs and outdoor BSs, even for antennas in proximity, due to the determining influence of the environment type on the indoor user activities. This phenomenon has not been highlighted or quantitatively analyzed before.
- We reveal that the total Internet traffic data as well as the traffic generated by the individual applications exhibit different activity peaks and temporal patterns in the various identified clusters. That paves the way for the proactive management of ICN traffic by mobile traffic operators (MNOs).
- To make our work reproducible and allow other researchers to benefit from our findings, we will make publicly available the code and processed service consumption data described in Section 4.1 that we use to conduct our analysis.

2 RELATED WORK

Internet measurement data produced by mobile communication systems has been used across various research fields such as sociology, ecology, epidemiology, economics, transportation, and telecommunication & network engineering [32, 10, 25, 47, 40]. These studies aim at educing conclusions by observing the generated traffic and leveraging it as a means to comprehend and interpret human behavior and drives. For instance, Internet service usage was exploited to assess the socioeconomic status of individuals [43] or to portray the population behavior during different pandemic phases [56].

Works stemming from the field of telecommunication and network engineering focus more on the traffic characteristics *per se*, rather than correlating them to socioeconomic, climate, or epidemic features. In [17], exploratory factor analysis was used for network activity profiling and to identify time periods that yield an unvarying mobile traffic demand spatial distribution. To characterize temporal data usage patterns, a two-state Markov model was proposed in [28], and it was shown that two distinct user groups exist, having unique usage and service preferences. More recently, deep learning (DL)-based approaches were explored; building on a clustering algorithm employed to dissever the city areas into different clusters, a DL architecture was exploited in [52] to predict the spatio-temporal cellular traffic. A graph neural network-based approach was presented in [57], for the joint spatio-temporal analysis and prediction of traffic demands, considering separately in-cell and inter-cell traffic. However, all these approaches build upon the cellular data recorded at outdoor BSs, disregarding the ICN traffic and its peculiarities.

The literature on the characteristics of ICN traffic demands is in fact rather limited [44, 60, 24]. The works in [44, 60] compared the characteristics of outdoor BSs and wireline traffic demands, pronouncing that there are distinct differences in terms of packet, flow, and session-level statistics, as well as in the temporal traffic patterns. However, despite wireline networks being installed indoors, their operation is fundamentally different from that of wireless ICNs, and thus the conclusions of [44, 60] cannot be straightforwardly extended to our case study. The traffic from ICNs was considered in [24], exploring the mobile application utilization profiles in Santiago de Chile for a 15-day period. The authors discussed how the urban context and time of the day affect human activities and application usage. More specifically, along with the application traffic patterns of outdoor BSs, they presented these of indoor BSs, but without expanding on their specific characteristics or how they fluctuate across different indoor environments. Our work fills that gap, building upon a rich Internet measurement data set to evince

that various application utilization clusters exist inherently in ICN traffic, revealing the particularities of each cluster, and correlating them with the indoor environment in which they are generated.

3 INTERNET MEASUREMENT DATA

The measurement traffic data used in our work was collected in the production network of a major MNO providing services throughout a nationwide network in France. To conduct the measurement collection, the operator deploys passive measurement probes to monitor the Gi, SGi, and Gn interfaces that connect Gateway GPRS Support Nodes (GGSNs) and Packet Data Network Gateway (PGWs) to external Public Data Networks (PDNs), gathering data about the traffic generated by mobile subscribers using 4G and 5G connectivity. Indeed, the MNO is presently operating a 5G non-standalone (NSA) deployment, which allows gathering data about both 4G and 5G traffic via measurements limited to the 4G Evolved Packet Core (EPC) network that is shared by 4G eNodeBs and 5G gNodeBs in the Radio Access Network (RAN).

The MNO identifies the mobile service associated with each TCP and UDP session recorded by the probes, by running Deep Packet Inspection (PDI) and analyzing the results via proprietary traffic classifiers. Each such IP session is also geo-referenced at the level of Base Transceiver Station (BTS), by exploiting the User Location Information (ULI) field present in the Packet Data Protocol (PDP) Contexts and Evolved Packet System (EPS) Bearers over the GPRS Tunneling Protocol control plane (GTP-C). Based on this information, it is possible to estimate the traffic demand for each mobile application at an individual BTS, by aggregating the traffic of all IP sessions from/to the same BTS. Such data is aggregated over time within intervals of one hour for the purpose of our study.

Overall, the data used throughout the paper comprise per-hour downlink and uplink traffic of various mobile services at all indoor cellular antennas managed by the MNO in the target country. This includes 4, 762 ICN antennas installed at more than 1, 000 sites, comprising different types of indoor environments located at an urban, suburban, or rural surrounding. We note that the vast majority of those antennas are 4G, as apparently 5G is scarcely used for ICN at this stage of roll-out of the technology in France. The recording period was approximately two months, starting from the 21$^{\text{st}}$ of November 2022 and extending until the 24$^{\text{th}}$ of January 2023. The mobile services considered span a diverse range of mobile applications used throughout daily life related to activities such as social networking, messaging, audio and video streaming, transportation, professional activities, and well-being. The ethical considerations related to the data collection and processing are discussed in detail in the Ethics sections in the Appendix.

4 CLASSIFYING ICN BEHAVIORS

In this section, we present the methodology followed to study the dynamics of ICN traffic and unveil the distinct patterns that reside in the data set described in Section 3.

4.1 Measuring relative ICN service usage

To probe the dynamics of the ICN mobile traffic demands across the network, we pursue an unsupervised learning approach. In particular, we consider the aggregated sum of the downlink and uplink traffic recorded for each mobile service over the target two-month period as a distinct feature, and we form a matrix, $T^{N\times M}$, comprising the overall traffic in megabytes (MB) for the N indoor antennas and M applications included in our data set. In particular, for the rest of our study, we consider $N = 4,762$ indoor antennas and $M = 73$ mobile services.

Our aim is to pass this matrix to an unsupervised learning algorithm, which will be able to identify patterns and existing hidden structures within the traffic data, without further guidance. However, the overall traffic is not a defining feature *per se*, and it will inevitably instigate bias in the clustering algorithm for two reasons. First, some applications intrinsically produce a larger volume of traffic than others, *e.g.*, streaming services generate demands that can be orders of magnitude larger compared to those induced by texting applications. Second, due to their location, antennas are also expected to serve highly heterogeneous traffic volumes. Therefore, clustering directly the aggregated traffic would result in overlooking the impact of many services and essentially grouping together antennas according to their popularity. This is in opposition to our goal to quantify the diversity of application usage across the various antennas.

To overcome these limitations, we consider a more representative and fair measure, *i.e.*, the revealed comparative advantage (RCA), which was initially employed in the field of international economics [7]. The RCA is an index of the relative advantage or disadvantage of a specific sample in a certain category. In our case, RCA quantifies the degree of over- or under-utilization of a certain service at a specific antenna. Given the matrix T, the RCA per application for each antenna can be computed as:

$$RCA_{i,j} = \frac{T_{i,j}/T_i}{T_j/T_{tot}}, \tag{1}$$

where $T_{i,j}$ stands for the traffic recorded for the j-th service at the i-th antenna, T_i refers to the total traffic generated at the i-th antenna for all the services, T_j depicts the summed traffic over all antennas for the j-th service, and T_{tot} is the total traffic channeled through the network during the entire data collection period.

Values of RCA below 1 indicate that an antenna is disadvantaged in the use of a certain service with respect to the other antennas, *i.e.*, there is an under-utilization of the service by that antenna, while values higher than 1 suggest that an antenna is advantaged, *i.e.*, there is an over-utilization for the same service. Evidently, from the definition of RCA it follows that while under-utilization is bounded to 0, over-utilization is unbounded and ranges from 1 to infinity. That is a well-known shortcoming of RCA that can lead to misleading results and curb the performance of the unsupervised learning clustering algorithm. A remedy for that was discussed in [29, 30], by making the index symmetric, thus enabling an equitable comparison between the under- and over-utilization intervals. In particular, Laursen and Engedal introduced the revealed symmetric comparative advantage (RSCA) which is defined as follows [29]:

$$RSCA_{i,j} = \frac{RCA_{i,j} - 1}{RCA_{i,j} + 1}, \tag{2}$$

from which it emerges that the values of RSCA lie within the $[-1, 1]$ interval, with values below 0 yielding under-utilization and values higher than 0 implying over-utilization.

Figure 1: Histogram of the normalized traffic, the RCA, and the RSCA for the M mobile services for some antennas.

Figure 1 shows the histogram of the (i) normalized traffic (over the maximum application load observed among all selected antennas), (ii) RCA, and (iii) RSCA of the M mobile services for some antennas of the data set. As it is normalized by the application with the highest load observed at any antenna, the normalized traffic presents a spike-like behavior with most applications having values squeezed close to 0, due to their inherently lower produced load or infrequent use, as well as owing to the less central location of the antenna. A closer look can be obtained when zooming between the 0 and 0.5 interval (upper right corner of Figure 1), where it can be seen that the number of popular mobile services with higher load drops extremely quickly. Ultimately, the presence of highly imbalanced demands across antennas and services makes a straightforward normalization by the global maximum not viable, as it tends to hide the preponderant low-traffic samples.

On the other hand, the RCA values are clearly better (*i.e.*, more diversely) distributed for the same set of samples; however, one can observe that the distribution is skewed, with the underutilized services wedged between 0 and 1, whereas the over-utilized services RCA values span from 1 to beyond 5 (specifically, the largest RCA in this example is 75.88). Thus, RCA still includes outliers, designated by the histogram tail, which are precisely the points that can jeopardize the reliability of the clustering by drawing the multidimensional cluster baricenters towards cases with higher-than-average usages. Such a behavior is removed from the data in the RSCA distribution, which yields a properly balanced distribution among samples that show lower- and higher-than-usual application consumption. Hence, in the impeding clustering analysis, we select the RSCA of each application as a distinct feature, separating the data based on the utilization profile of the M mobile services across the N distinct antennas.

4.2 Clustering ICN antennas

Having defined a proper transformation of the recorded traffic, the next step is to feed the processed data to an unsupervised learning algorithm. The aim is identifying clusters of ICN antennas based on similarities and differences in the considered set of futures, *i.e.*, the RSCA values that denote under- or over-utilization of the M applications at each antenna with respect to typical usages.

Figure 2: Silhouette score and Dunn index versus the number of clusters, serving as a stopping criterion to select the optimal number of clusters.

4.2.1 Clustering strategy. While multiple techniques are available for unsupervised learning clustering, due to its comprehensibility we opt for agglomerative clustering, which is a state-of-the-art hierarchical clustering algorithm [39]. The algorithm follows a bottom-up approach, assuming initially that each data point belongs in its own cluster, and thereafter merging the clusters greedily according to a specified criterion. In particular, we use Ward's criterion [53] which aims at minimizing the total intra-cluster variance, measured as the squared distance between the cluster centers, when merging two clusters. Hence, the clustering algorithm starts from N distinct clusters and repeatedly merges the clusters, reducing their number in a way that the new cluster yields a reduced intra-cluster variance.

During this process, an important point is to determine when the merging should stop, *i.e.*, which is the number of the total clusters, k, that separates optimally the data. To quantify the performance of the clustering algorithm and determine the most suitable number of clusters, we employ the Silhouette score and the Dunn index [45, 13]. The first expresses the degree of similarity among objects of the same cluster (cohesion) compared to other clusters (separation), while the latter indicates how compact (small inter-cluster variance) and well-separated (large intra-cluster distance) the clusters are. To identify the most suitable number of clusters, we seek a high value of the Silhouette score or the Dunn index, followed by an abrupt drop, which suggests a substantial deterioration of the intra- and inter-clustering quality. As can be seen from Figure 2, such a behavior can be observed for $k = 6$ and $k = 9$. Hence, in what follows we select $k = 9$, which exhibits the steepest drop for both metrics, while we will also discuss qualitatively the differences that occur for a reduced number of clusters $k = 6$.

4.2.2 Clustering results. The clustering results based on the mobile service RSCA are summarized in Figures 3 and 4. Figure 3 depicts the complete hierarchy returned by the agglomerative algorithm on the service SRCA values, starting from the individual antennas at the bottom and reaching a single cluster at the top. The optimal clustering identified by $k = 9$ is also highlighted, with the set and number of antennas in each final cluster reported along the x-axis of the dendrogram. In Figure 4, we instead present a complementary RSCA heatmap, clumping the antennas per cluster, and presenting

Figure 3: Dendrogram illustrating the iterative merging of antennas into clusters as returned by the hierarchical agglomerative algorithm run on SRCA features of individual antennas. Distance thresholds for $k = 6$ and $k = 9$ are highlighted. Colors tell apart the 9 clusters identified by the second threshold.

the RSCA of all the indoor antennas for each service. Blue lines in the heatmap suggest over-utilization of a certain service across the antennas, while the opposite holds for dark red lines. Evidently, it can be observed that RSCA follows a distinct visual pattern for each cluster, *i.e.*, the indoor antennas appertaining to the same cluster manifest the same pattern with respect to service utilization. The RSCA pattern can be substantially different compared to that of the other clusters, but there are also some clusters between which the differences are soothed.

To acquire further insight regarding the degree of similarity between the 9 clusters, one can peruse the hierarchical algorithm dendrogram in Figure 3, in which we draw two horizontal lines to indicate the distance threshold that distinguishes the clusters for $k = 6$ and 9. From the dendrogram, and considering that dissimilarity in terms of service usage is associated with distances along the y-axis, it emerges that there are three large groups of clusters (ending at the edge of the blue lines), which are further separated into three sub-clusters colored orange, green, and red.

The first larger group comprises clusters 0, 7, and 4 (orange group), the second clusters 5, 6, and 8 (green group), whilst the third includes clusters 3, 1, and 2 (red group). The clusters found within the same group present a stronger similarity with each other, which is also corroborated by the RSCA heatmap in Figure 4. Instead, they demonstrate unalike mobile service utilization patterns compared to the clusters of the other two groups. Out of the three groups, the orange one presents the more unique behavior, being further away both from the green and red groups. Furthermore, within each group, clusters found under the same branch yield a higher resemblance to each other compared to the clusters found in a separate branch, *e.g.*, clusters 1 and 2 for the red group. Finally, it should be noted that applying the hierarchical clustering with $k = 6$, corresponds to consolidating the clusters of the orange group into a single cluster, instead of diving it into 3 sub-clusters, and merging clusters 6 and 8 at the second branch of the green group.

Although the dendrogram allows identifying groups of clusters that are closer to each other, it cannot be used to delve into the feature importance of each cluster. Hence, an intriguing question that arises from our analysis is: *which are the features, i.e., services, that impact the clustering decision the most, and consequently characterize each cluster?* To answer this question, we next explore which are the user tendencies and mobile service utilization patterns within each cluster that render them unique.

Figure 4: Heatmap of the mobile service RSCA (y-axis) for the clustered ICN antennas (x-axis), when $k = 9$.

4.2.3 Key insights. RSCA enables acquiring an unbiased representation of the mobile service utilization at ICN antennas. Clustering the antennas based on their RSCA yields 9 distinct service usage clusters that can be aggregated into 3 larger groups. As expected, clusters within the same group clearly demonstrate more similarities compared to clusters in other groups.

5 UNDERSTANDING ICN TRAFFIC

We now show how explainable ML techniques can be leveraged to interpret the results of our analysis and breakdown the application trends per cluster, and we correlate these trends with the type of indoor environment in which they occur. Later, we also provide a comparison between the traffic behavior recorded by the indoor antennas and that of the neighboring outdoor antennas, demonstrating the innate particularities of ICN traffic demands.

5.1 Interpreting Clustering Results

In order to interpret the results of the clustering of ICN antennas, we first introduce the SHAP framework and explain how it can be applied to our use case. Then, we discuss its results and show how it helps to understand the under- and over-utilization of mobile services in each cluster.

5.1.1 The SHAP Framework. Although ML models have found application in various tasks, a major criticism against them is the lack of intuition behind their decisions. Indeed, their decisions are a result of consecutive complex non-linear operations applied to some input data, which renders their comprehension by humans extremely difficult or, in most cases, impossible. That has motivated a substantial research effort to create frameworks that enable the interpretation of ML models [22]. Among the various approaches, the SHAP framework has shown to be advantageous with respect to differentiating among the various output classes, but more importantly, also with respect to conforming better with human intuition when compared to other frameworks [31].

In particular, the SHAP framework aims at decomposing the response of an ML model to a set of linear explanation polynomial functions with binary random variables. Formally, let u be a polynomial explanation function, then [33]:

$$u(x') = \phi_o + \sum_{i=1}^{M} \phi_i x'_i, \tag{3}$$

where $x' \in \{0,1\}^M$ represent the binary random variables which are mapped to initial input space via a function q, *i.e.*, $x = q(x')$, and as defined previously, M is the number of the input features. The goal is to select the coefficients, ϕ, of the explanation function such as to approximate locally the black-box response of the ML model, *i.e.*, $u(x') \approx f(q(x'))$. Then, for a specific input, x', due to the simplicity of the linear model, the coefficients quantify the contribution of each feature towards the output. Indeed, from (3) it follows that a positive coefficient associated with a high feature value contributes positively to decision-making, whereas the opposite stands for a negative coefficient. Furthermore, the contribution of a feature diminishes as the absolute coefficient value decreases. As discussed in [33], the explanation functions u must satisfy a local accuracy, a missingness, and a consistency property, and to abide by these criteria, it was shown that there is a unique u, for which the polynomial coefficients are equal to the Shapley values, defined as [48]:

$$\phi_i(f, x) = \sum_{z' \subset x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f(z') - f(z' \backslash i) \right], \tag{4}$$

where $z' \subset x'$ indicates the possible vectors z' whose non-zero entries constitute a subset of the binary random variable entries x'. The notation $|\cdot|$ is used to represent the number of non-zero entries in a vector, while the notation $z' \backslash i$ implies setting $z' = 0$ for the i-th feature while keeping intact the remaining values of z'. The difference between the output of the black-box ML model with and without considering the i-th feature, *i.e.*, $f(z') - f(z' \backslash i)$, expresses the contribution of the feature in the model's response. Finally, the term multiplying the difference is used to weigh each contribution based on the number of non-zero features included in z'. During the Shapley value estimation, all the possible permutations in x' are considered, in order to probe the sole impact of a feature, as well as how the feature interplays with the other features to influence the output of the model.

An important remark regarding the estimation of the Shapley values via (4) is that it requires evaluating the model's response while removing the i-th feature from the input tensor, i.e, evaluating $f(z' \backslash i)$. However, typically ML models use constant-size input tensors, and thus it is not possible to remove entirely an input feature. A remedy to that is to replace the value of the feature that needs to be excluded with a peer feature random value coming from the training data set. By applying that ploy for all the possible vectors z', one can anticipate that the impact of the feature sought to be removed will be ultimately sampled out, and thus it is implicitly removed. Then, the estimation of the Shapley values can be performed via model-agnostic approximations, such as the Kernel SHAP that can be used to interpret any ML model [33], or through model-specific approximations, *e.g.*, TreeShap that is employed for tree-based ML algorithms such as random forests or XGBoost [34]. The model-specific approximations are commonly dramatically faster compared to the model-agnostic implementations, especially for large data sets and high-dimensional input spaces, as they are designed to explicitly optimize the Shapley value computation process for a specific ML architecture.

5.1.2 Interpreting clustering with SHAP. In this subsection, we leveraged the SHAP framework, aiming at interpreting the agglomerative clustering decisions presented in Section 4.2 and shedding light on the traffic demands of each cluster. The agglomerative algorithm forms the clusters by gradually linking samples together based on their distance from all the other samples, *i.e.*, the clustering is the result of consecutive iterations over the entire data set. It is important to note that the clustering is performed considering concurrently all the samples: therefore, the result cannot be used to infer the cluster of a single sample from the data set by solely observing its input features, or to predict the cluster of new samples not included in the data set. This implies that once the clustering is completed, there is no black-box function f to be interpreted.

A common approach to overcome this limitation is to build a surrogate supervised learning classifier on top of the unsupervised learning clustering results, treating the clusters as the target labels associated with each sample of the supervised learning problem [38, 16, 8, 21, 3]. The supervised learner is typically a decision trees-based model, and once it is trained to approximate the clustering results, one can instead interpret the predictions of the supervised learning classifier. Thence, considering as ground truth labels the clusters of Section 4.2, we train a random forest classifier with 100 trees to infer the antenna cluster based on the mobile service RSCA, and we then amalgamate it with the SHAP framework presented earlier. That also allows exploiting the expediency of the TreeShap method for the estimation of the Shapley values [34].

With the Shapley coefficients evaluated, one can inspect their values to comprehend the importance of each service in the cluster decision-making process, as well as whether the importance springs from the under- or over-utilization of the service. The computed coefficients for each cluster are presented as beeswarm plots in Figs 5a to 5i, where the red and blue color implies that the coefficient is associated with higher and lower feature values, respectively. A positive SHAP value associated with a high feature value signifies a positive impact on the prediction, leading the model to categorize the antenna in the respective cluster due to the corresponding

(a) Orange group, cluster 0. (b) Orange group, cluster 4. (c) Orange group, cluster 7.

(d) Green group, cluster 5. (e) Green group, cluster 6. (f) Green group, cluster 8.

(g) Red group, cluster 1. (h) Red group, cluster 2. (i) Red group, cluster 3.

Figure 5: SHAP values beeswarm plots for clusters 0 to 8, depicting the importance of each application in the clustering decision. The group color refers to the branch color of the dendrogram in Figure 3. Positive Shapley values matched with high feature values imply that for an antenna to belong to that cluster it should be overutilizing the respective application, while low feature values with positive coefficients suggest under-utilization.

application over-utilization. Conversely, positive coefficients, combined with low feature values, indicate that the antenna belongs to a cluster characterized by under-utilization of a certain application.

To facilitate the analysis and the comprehension of the clustering criteria, the results are presented according to the groups identified in the dendrogram of Figure 3. The Internet services in each beeswarm plot are ranked in descending order according to their impact on the model's output. The impact of each service is measured as the mean absolute value of the service coefficients, *i.e.*, applications with high coefficient values influence cluster inference more. For each cluster, we present the 25 most influential services.

Remarkably, we can observe that there are similarities in the mobile service utilization patterns for the clusters falling under the same color groups discussed in Section 4.2. For instance, the antennas of the orange group share in common that they over-utilize applications related to music, such as Spotify, Soundcloud, Deezer,

Table 1: Summary of Indoor environment types.

	Metro	Trains	Airports	Workspaces	Commercial Centers	Stadiums	Expo Centers	Hotels	Hospitals	Tunnels	Public Buildings
Cases	Paris, Lille, Lyon, Rennes & Toulouse underground railways	National & regional railway stations	France's major airways	Corporate offices Industrial facilities	Malls Shopping Stores	Major sport event venues	Corporate, cultural & music event venues	Accommo-dation units	Healthcare units	Highway & train tunnels	Universities Museums
N_{env}	1794	434	187	774	469	451	230	28	53	220	122

and Apple Music. Likewise, applications related to Navigation, such as transportation websites, Mappy, and Google Maps, have a strong presence in clusters 0 and 4, and that in fact distinguishes them from cluster 7, which is characterized by under-utilization of these applications. Finally, unlike cluster 0, the utilization of entertainment services is scarce in cluster 4, *e.g.*, Yahoo and entertainment, shopping, or sports websites, which explains their separation into different clusters.

On the other hand, the main similarity between the green group clusters is that they exhibit strong under-utilization for most of the mobile services. That is more apparent in cluster 5, which constitutes a distinct leaf in the dendrogram, whereas clusters 6 and 8 both belong in another branch. Moreover, both clusters 6 and 8 include antennas that over-utilize Snapchat, Twitter, and Sport websites. However, cluster 8 exhibits a larger diversity of applications, which renders it unique: services such as Giphy, WhatsApp, and streaming such as Canal+ are absent in cluster 6.

The red group exhibits a substantially different behavior compared to the orange group, as clusters 1, 2, and 3 demonstrate minor utilization of music and navigation-related applications, including Spotify, SoundCloud, Mappy and Transportation websites. Moreover, cluster 3, which is a distinct leaf in the dendrogram, deviates from clusters 1 and 2, since it is characterized by intense use of business-oriented applications, such as Microsoft Teams, LinkedIn, and emailing services. On the other hand, cluster 1 presents over-utilization of diverse application types such as streaming services (Netflix, Disney+, Amazon Prime Video), vehicular navigation (Waze), and mailing applications. Finally, cluster 2 is characterized by services related to digital distribution services (Google Play Store) and Shopping websites.

5.1.3 Key insights. The application utilization analysis corroborates the sanity of the unsupervised learning approach followed in Section 4.2. Indeed, it is now evident that clusters found within a small distance, *i.e.*, in the same group, yield small differences in their mobile traffic demand patterns, whilst the trends among clusters of different groups are substantially different. Despite that though, an intuitive explanation and the roots of this phenomenon are still unidentified.

5.2 Indoor Environment Analysis

The data provided by the MNO is accompanied by additional information related to the antennas, such as their geographical location, the location area code, and the name of the BS. We next exploit such information to foster the interpretability of the clustering results.

5.2.1 Indoor environment types. By inspecting the names of the antennas, applying simple string manipulation to extract keywords appearing within the names, and cross-checking the respective antenna locations online, we are able to identify specific types of indoor environments in which ICNs are deployed. We individuate the following eleven categories of indoor locations: (i) metro stations, (ii) train stations, (iii) airports, (iv) workspaces, (v) commercial centers and shopping stores, (vi) stadiums, (vii) expo centers, (viii) hotels, (ix) hospitals, (x) tunnels, and (xi) public buildings.

A summary of the indoor environment types identified can be found in Table 1, along with the number, N_{env}, of the antennas per environment. As can be seen from Table 1, the distribution of indoor environment type is unbalanced, which is expected since some environments are encountered more often than others across cities, *e.g.*, there are more offices than hospitals, and the MNOs deploy ICNs in specific locations according to their interest. It should be noted that this imbalance in the data does not impact our analysis; indeed, we do not pursue a supervised learning approach (where one has predefined class labels and should be careful in the training of the model to account for the imbalanced classes), rather we inspect the problem from the angle of unsupervised learning. This is to say, that the algorithm splits the data without guidance, solely based on the underlying service utilization patterns it identifies, and without prior knowledge of the antenna environment or any other antenna-related information. Thus, it's reasonable to assume that if there were more antennas with the same mobile service utilization profiles, they would have been distributed accordingly.

The metro stations cover the entire underground transportation system of metropolitan Paris as well as the Parisian RER (Réseau Express Régional) stations, whilst there are also metro BS in the cities of Lille, Rennes, Toulouse, and Lyon. Likewise, the train stations constitute the departing and arriving points of the French railway system. The airports include Paris main flying arteries, *i.e.*, Charles de Gaulle and Orly airports, and smaller aerodrome units found in multiple other cities, such as Bordeaux and Nantes. Workspaces comprise corporate offices, headquarters, and a limited number of industrial facilities of major companies operating in different domains, *e.g.*, Ericsson, HSBC, AXA, Givenchy, and Lacoste. Commercial places consist of malls, shopping centers, department stores, home furnishing outlets, large supermarkets, and the small retail shops of the MNO where numerous indoor antennas are also installed. The stadiums correspond to facilities that host major sports games and tournaments in France, such as Stade de France, while the expo centers house various corporate, cultural, and music events. The tunnel antennas in our data set provide coverage to vehicle or train on-board users entering tunnels, *e.g.*, underground tunnels or tunnels crossing mountains. Finally, the public building category includes administration buildings, universities, schools, and cultural heritage buildings throughout France.

5.2.2 Interpreting ICN clusters via indoor environments. By leveraging the environment information above, we are able to associate classes of antennas with their specific surrounding conditions. A

Figure 6: Sankey diagram depicting how the clusters flow into different environment types.

logical next step is then to quantify the type of indoor environments that reside within each cluster. A qualitative illustration of the correlation between the detected clusters and the indoor environment type is shown in the Shaneky diagram of Figure 6, depicting how the samples of each cluster flow into the environments of Table 1. Conspicuously, there are intense streams between certain clusters and environment types. For instance, it can be seen that the metro and train stations are monopolized by the orange group clusters, while the preponderance of stadiums appertains within one of the green group clusters. Furthermore, it can be observed that the dominant flux towards workspaces originates from cluster 3, whilst clusters 1 and 2 populate the remaining environments.

A quantitative analysis is presented in Figure 7, where we illustrate the percentage of each cluster occupied by the different environment types. In addition, in Figure 8, we report how the different environment types are distributed among the various clusters for some non-obvious cases.

As mentioned, a striking observation is that the orange group clusters (0, 4, and 7), shown in Figure 7a, comprise solely metro and train stations. That further rationalizes the mobile service utilization patterns discussed in Section 5.1.2, and correlates them with the habits and the activities in which the users are engaged in relation to the environment in which they are found. Indeed, the orange group represents antennas that serve users while commuting, which explains why they are likely to listen to music, use navigation applications to make their way or check public transportation schedules, and be involved with entertainment-related applications while traveling. Notably, for clusters 0 and 4 more than 92% of the antennas are located in Paris and its suburbs, contrary to cluster 7 which consists solely of the Lille, Lyon, Rennes, and Toulouse metro antennas, *i.e.*, non-capital cities.

That further clarifies the distinction between the antennas of the orange group: non-popular navigation applications, such as Mappy and Transportation websites, are likely to fall in under-utilization when compared to the usage patterns of metropolitan commuters. Indeed, the RSCA does relate to the absolute volume of data, rather it captures the peer-service usage between different antennas. Moreover, the comparative under-utilization can be also attributed to the complexity of the Parisian metro system compared

(a) Orange group.

(b) Green group.

(c) Red group.

Figure 7: Types of indoor environments per cluster.

to that of smaller cities, *i.e.*, a single route might entail multiple station and line changes that are more complex to navigate and need support of dedicated apps. Finally, the fact that Paris is a touristic destination fosters the use of navigation services by non-resident individuals.

Another interesting remark is that the preponderance of the antennas found in the green group is stadiums, as can be seen from Figures 7b and 8c. That justifies the high number of visits to sports websites and the use of content-sharing applications,

(a) Airports, tunnels, commercial centers. **(b) Hotels, hospitals, and public buildings.** **(c) Stadiums, expo centers, and workplaces.**

Figure 8: Cluster distributions per indoor environment type.

such as Twitter and Snapchat via which one can upload photos and information relevant to sports events. This behavior is more evident for the antennas in clusters 6 and 8, more than 75% of which are in stadiums. Remarkably, cluster 6 includes stadiums outside Paris, while approximately 60% of cluster 8 antennas are in Paris.

On the other hand, stadiums make up only 35% of cluster 5, which also includes other diverse types of environments, such as expo centers, corporate offices, and commercial centers, equally distributed in Paris and other cities. Apart from the expo centers though, the remaining categories represent only a small fraction of their environment type, *e.g.*, only approximately 5% of the commercial centers and working environments belong to cluster 5, as can be seen from Figures 8a and 8c, respectively. Hence, given that cluster 5 is characterized by the under-utilization of most mobile services, we consider it to include antennas treating most of their Internet services equally, without demonstrating any advantages compared to other antennas in the data set. In other words, service usage is equally distributed at those antennas, yielding a similar small numerator for all services in (1), compared to a larger denominator.

Expatiating on the environments encountered in the red group, the most apparent connection is observed for cluster 3. As can be seen, more than 70% of cluster 3 antennas are workplaces, in particular corporate offices, which clearly provokes a surge in the use of services such as Microsoft Teams, LinkedIn, and emailing applications. Notably, in contrast with the offices, the small number of industrial facilities included in the data set mostly occupies cluster 5. Expo centers also assume a strong presence, with more than 50% of them belonging to cluster 3 as seen in Figure 8c, likely due to holding corporate events, conventions, and conferences.

Interestingly, while cluster 3 constitutes a single branch of the red group, clusters 1 and 2, which belong to the other branch, host many commercial centers. In particular, cluster 2 hosts 50% of the commercial centers, most of the hotels and public buildings, as well as almost all the hospitals, as shown in Figure 8b. Notably, cluster 2 includes all the small retail shops of the MNO, which clarifies its advantage in Google Play Store usage to download apps.

Cluster 1 is more diverse, and along with commercial centers, it also contains almost all airport and tunnel antennas, as illustrated in Figure 8a, and a small percentage of all the available environment types. Given the large variety of environments encountered in cluster 1 and the absence of a dominant environment type, as well as due to the fact that it contains messaging, streaming, and music services, it can be speculated that this serves as a general-use cluster. As a final note, we mention that while the antennas of cluster 1 are almost equally distributed between Paris and other cities, at around 92% of the antennas of cluster 2 are found outside Paris, and 70% of cluster 3 antennas are located in Paris and its suburbs.

5.2.3 Key insights. Explainable machine learning allows comprehending the clustering criteria. Consequently, we discovered that distinct clusters were characterized by the use of music and navigation applications, general-purpose applications, and work-oriented or sport-related mobile services. Observing the most and least important mobile services per cluster and juxtaposing them to the indoor environment, we pronounce that they are substantially influenced by the latter.

5.3 Comparison with Outdoor Antennas

From the analysis presented earlier, it clearly emerges that ICN mobile traffic demands intrinsically exhibit distinct patterns that highly depend on the type of indoor environment. A natural question is whether these patterns are also present in the components of legacy communication system radio access networks. In order to answer that, we probe the traffic generated by outdoor antennas found in proximity to the ICNs of our data set.

5.3.1 Quantifying outdoor service usages. In particular, we want to assess whether the nature of the ICN strictly defines and differentiates its service demands from that of close-by outdoor antennas, *e.g.*, if the service demands of the outdoor antennas nearby to corporate offices are drastically disparate from the in-building ones. Hence, for each indoor antenna, we consider all the outdoor antennas found within a 1km radius, and we compute their RCA as:

$$RCA_{out_{i,j}} = \frac{T_{out_{i,j}}/T_{out_i}}{T_{in_j}/T_{tot,in}}, \tag{5}$$

where $T_{out_{i,j}}$ represents the traffic recorded for the j-th service at the i-th neighboring outdoor BS, T_{out_i} is the total traffic generated at the i-th outdoor BS for all the mobile services, while the ratio $T_{in_j}/T_{tot,in}$ expresses the level of utilization the of j-th service over the entire indoor traffic. Then, the RSCA for the outdoor antennas can be computed straightforwardly via (2).

It should be noticed that according to (5) the RCA for the outdoor antennas measures the level of utilization for a certain service in an outdoor antenna compared to the level of the same service usage

Figure 9: Distribution among the identified clusters, for 22,000 outdoor antennas located in close proximity of the ICN antennas considered in our study.

among all the indoor antennas. Indeed, rather than studying the traffic observed at outdoor antennas per se, we want to explore whether this traffic is innately different compared to that generated in indoor environments.

5.3.2 Comparing indoor and outdoor demands. With the RSCA for all the neighboring outdoor antennas estimated, their cluster can be inferred by feeding these values to the trained random forest classifier used in Section 5.1.2 to surrogate and generalize the unsupervised learning clustering results. The predicted cluster distribution for approximately 20,000 neighboring outdoor antennas is presented in Figure 9. Evidently, the innate diversity encountered in the traffic demands of indoor antennas is absent for outdoor antennas. Indeed, almost 70% of the outdoor antennas appertains in cluster 1. That further corroborates our intuition that cluster 1 constitutes a general-use cluster. More importantly, though, it becomes clear that the distinct traffic behavior of workplaces, stadiums, metro, and train stations is now almost absent, as only a negligible percentage of the outdoor antennas lies within the respective clusters.

5.3.3 Key insights. The intrinsic service demand diversity observed in ICNs is absent from the neighboring outdoor BSs, despite their close proximity. This behavior bespeaks that ICN traffic is highly environment-centric, whereas outdoor BSs accommodate general-purpose traffic.

6 TEMPORAL ANALYSIS

The next step in our analysis is to uncover the patterns that the ICN traffic exhibits over time. Expectedly, one should observe different patterns for each cluster, as it has been highlighted in previous research which focused though solely on the traffic recorded at outdoor antennas [18, 17].

6.0.1 Cluster-level temporal demands. To this end, in Figure 10 we provide heatmaps showing the evolution of the normalized median traffic per hour across all the antennas belonging to the same cluster, for the period of 04/01/2023 to 24/01/2023. Evidently, it can be pointed out that there is a strong correlation between the temporal patterns and the indoor environment type. Indeed, as can be seen in Figure 10a to 10c, the orange group clusters, which are populated with metro and train stations, demonstrate a traffic peak during the common weekly commuting hours in France, *i.e.*, 7.30 to 9.30 a.m. and 17.30 to 19.30 p.m. In the remaining hours of the day the traffic volume is considerably smaller, while the same holds throughout the weekends, e.g., the 7^{th} and 8^{th}, or the 14^{th} and 15^{th} of January. Remarkably, there is another day with negligible traffic in the period under consideration; the 19^{th} of January, which corresponds to a national general strike day. The strike's impact is not as severe for cluster 7, which includes metros antennas found in cities other than Paris, presumably due to the milder impact of the strike in these cities.

The green group clusters temporal patterns, presented in Figs. 10d to 10f, are characterized by sporadic, non-canonical bursts of data usage. Again, this is an expected behavior for stadiums and convention centers, where unlike the other indoor environments discussed, surges of mobile subscribers appear on the premises and generate traffic only when events are taking place. For instance, in cluster 8 a traffic outbreak was observed only on the evening of January 19^{th} in a cross-Atlantic NBA special event conducted at Accor Arena in Paris, while the continuous burst in cluster 5 between the 19^{th} and the 24^{th} of January is a corollary of the 4-day Sirha Lyon event that took place at the Eurexpo Lyon convention center. Another difference between the green group clusters is that cluster 5 records a low volume of traffic throughout the entire day. Indeed, cluster 5 does not include only stadiums but other environments as well, which as discussed in Section 5.1.2 they all share in common a moderate application usage, as indicated by the mild blue color.

As depicted in Figs 10g to 10i, the red group clusters present a more vivid and diurnal pattern compared to the other two groups, with the traffic being almost evenly distributed from 10 a.m. to 8 p.m. Interestingly, clusters 1 and 2, which comprise numerous commercial antennas, record traffic throughout the weekdays and weekends, while cluster 3 consisting primarily of workspaces remains idle during weekends and after working hours, *i.e.*, 5.30 pm. That is a unique behavior that clearly differentiates cluster 3 from the other two clusters. Delving further into the peculiarities of each cluster, cluster 2 yields a slight drop on Sundays, which can be attributed to the fact that it contains smaller stores, *e.g.*, the MNO agencies, which are not as active on those dates compared to larger commercial stores and malls. A further difference between clusters 1 and 2, is that the latter presents higher traffic during nighttime, likely due to the larger number of hotels and hospitals, as shown in Figure 8b, which are more active during these hours.

6.0.2 Service-level temporal demands. Finally, further insight can be gained by inspecting the temporal heatmaps for some mobile services that assumed a key role in the cluster decision-making process, as per Figure 5. Hence, for the orange group clusters we select Spotify, Twitter, and transport websites, depicting their normalized median traffic as heatmaps in Figures 11a-11c. As it can be seen, Spotify assumes a strong presence for the entire group, with alike motifs for all clusters, demonstrating its traffic peaks during the morning commuting hours. On the other hand, the usage of transport websites is scattered for cluster 7, while clusters 0 and 4 preserve a more lively commuting hour pattern. Moreover, as

(a) Orange group, cluster 0. **(b) Orange group, cluster 4.** **(c) Orange group, cluster 7.**

(d) Green group, cluster 5. **(e) Green group, cluster 6.** **(f) Green group, cluster 8.**

(g) Red group, cluster 1. **(h) Red group, cluster 2.** **(i) Red group, cluster 3.**

Figure 10: Normalized median traffic heatmaps per hour for a period between 04/01/2023 and 24/01/2023. Each heatmap represents the median traffic of all antennas that belong to the specified cluster at a specific hour and day, while the light gray dashed lines indicate weekends.

(a) Spotify, Orange group **(b) Twitter, Orange group** **(c) Transportation websites, Orange group**

(d) Netflix, Green group **(e) Waze, Green group** **(f) Snapchat, Green group**

(g) Microsoft Teams, Red group **(h) Netflix, Red group** **(i) Waze, Red group**

Figure 11: Heatmaps of per hour normalized median traffic between 04/01/2023 and 24/01/2023, for the antennas of each cluster and for a selected set of services chosen based on their importance according to the SHAP values presented in Figure 5. The light gray dashed lines indicate weekends.

indicated by Figure 5b, Twitter usage is comparatively mitigated for cluster 4, since both clusters 0 and 7 present a persistent peak during morning or evening commuting hours.

The traffic patterns of the green group clusters, comprising event-oriented venues, are also in alignment with the conclusions drawn based on the SHAP values. For instance, Snapchat in Figure 11f presents traffic patterns significantly similar to the ones seen for total traffic in Figure 10, indicating the substantial usage of social media apps throughout these events. Likewise, in Figure 11e, Waze which is used for driving navigation and obtaining live traffic maps and road alerts, also shows an interesting traffic pattern. In particular, for these antennas, Waze assumes its peak values a couple of hours after the peaks of total traffic, as well as these of social media apps, which suggest the usage of vehicular navigation apps to guide the event attendants back to their home destination. In addition, video streaming apps, such as Netflix, fall into under-utilization in such venues, even on specific peak days and hours, as can be seen in Figure 11d.

For the red group, we select three very characteristic applications to highlight their different utilization during day hours; Microsoft Teams, Netflix, and Waze. Microsoft Teams, which is a work-oriented service, attains very small peak values for clusters 1 and 2 which are composed mostly of non-working environments. On the contrary, cluster 3 records heavy traffic over working hours, or even during the lunch break, since it is populated with workspaces. Meanwhile, Netflix exhibits the opposite pattern, yielding a stronger usage in daytime and nighttime for clusters 1 and 2, respectively, but only being utilized during lunch hours for class 3. Again, this behavior aligns with intuition since cluster 1 is of general use, cluster 2 includes most of the hotels where guests use streaming services during nighttime, while in working environments the use of streaming services is strictly limited during breaks. Finally, corroborating the findings of Figure 5g, out of the red group clusters, Waze attains the highest importance and recorded traffic for cluster 1, which aligns with intuition since as per Figure 8a it includes the majority of tunnels. Furthermore, the Waze traffic peaks in cluster 1 occur mostly on Saturdays, whereas for cluster 3 the heaviest traffic is recorded on weekdays after office hours when the employees return to their residences.

6.0.3 Key insights. The clusters designated by our analysis demonstrate diverse characteristics in the overall and per-application utilization patterns over time, which can be ascribed to the type of indoor environments associated with each cluster.

7 DISCUSSION AND ROADMAP

Our analysis unveiled that ICN mobile service demands are site-specific and more specialized than outdoor ones; therefore, ICN resource orchestration should not target overall capacity, as in outdoor environments, but must take into account the most important application usage per indoor environment. In a sense, our work fosters adopting a distinct network slicing dimension for indoor network resource planning, where the indoor slices will be tuned based on the characterizing applications for that specific indoor environment. Applications of such slicing could include adaptive power transmission control or content caching according to the insights provided by our analysis.

We expect the service usage trends identified in our analysis to remain timely for non-standalone 5G deployments. However, over time, with the emergence of applications such as the industrial Internet of Things, augmented reality, and intelligent self-orchestrated environments, we believe that additional clusters may emerge within ICN traffic, requiring further research and provisioning by MNOs. Indeed, Internet traffic produced by such applications will have much more specialized requirements and characteristics, hence we assume that this will intensify the environment and activity-oriented behavior underpinned by our analysis, giving further rise to the need for dedicated planning. Furthermore, as prior work has indicated [51, 42, 23], we believe that the results of our analysis will also be applicable to other countries that assume the same socio-economic standing and their residents are engaged in similar daily life activities. However, the impact of the socio-economic level of different countries on ICN traffic remains an open question for future research, as ICN traffic may not be affected by such factors due to its strong association with the type of indoor environment.

8 CONCLUSIONS

The proliferation of ICNs and the increasing mobile service demands generated by users in indoor environments call for an improved understanding of indoor traffic characteristics. Our work constitutes a primer in this direction, unveiling the unique behaviors that are intrinsic to the traffic generated by ICNs. We demonstrate that a set of clusters, with distinct mobile service utilization profiles, exists across a nationwide ICN deployment. Through explainable ML, we provide a detailed analysis of the services that tell profile apart, and consequently evince that ICN traffic demands strongly relate to the type of indoor environment. These profiles are more diverse compared to those encountered in the traffic generated by conventional outdoor BSs, and they exhibit distinctive overall and per-application utilization temporal patterns, influenced by the kind of activities that take place in the different indoor environments.

Overall, our work paves the road to the characterization of service-level mobile traffic demands in indoor environments, and offers a number of insights into patterns that were not observed or demonstrated quantitatively before.

ACKNOWLEDGMENTS

This work was supported by the European Commission through the Horizon 2020 Framework Program, H2020-MSCA-ITN-2019, MSCA-ITN-EID, under Grant 860239, BANYAN and the French National Research Agency, under Grant ANR-22-CE25-0016, CoCo5G project.

REFERENCES

[1] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. 2016. Next generation 5G wireless networks: a comprehensive survey. *IEEE Commun. Surveys Tuts.*, 18, 3, 1617–1655.

[2] Adnan Aijaz. 2020. Private 5g: the future of industrial wireless. *IEEE Ind. Electron. Magazine*, 14, 4, 136–145.

[3] Liat Antwarg, Ronnie Mindlin Miller, Bracha Shapira, and Lior Rokach. 2021. Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert systems with applications*, 186, 115736.

[4] Stefanos Bakirtzis, Jiming Chen, Kehai Qiu, Jie Zhang, and Ian Wassell. 2022. EM DeepRay: an expedient, generalizable and realistic data-driven indoor propagation model. *IEEE Trans. Antennas Propag.*, 70, 6, 4140–4154.

[5] Stefanos Bakirtzis, Marco Fiore, Ian Wassell, and Jie Zhang. Expedient ai-assisted indoor wireless network planning with data-driven propagation models. TechRxiv, (2023).

[6] Stefanos Bakirtzis, Ian Wassell, Marco Fiore, and Jie Zhang. 2022. Stochastic evaluation of indoor wireless network performance with data-driven propagation models. In *GLOBECOM 2022 IEEE Global Communications Conference*. IEEE, New York, NY, USA, 3587–3592.

[7] Bela Balassa. 1965. Trade liberalisation and "revealed" comparative advantage 1. *The manchester school*, 33, 2, 99–123.

[8] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. 2021. Interpretable clustering: an optimization approach. *Machine Learning*, 110, 89–138.

[9] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 6264, 1073–1076.

[10] W. Chen, Y. He, and S. Pan. 2021. Impact of air pollution on human activities: evidence from nine million mobile phone users. *PLoS ONE*, 16, e0251288, 5. DOI: https://doi.org/10.1371/journal.pone.0251288.

[11] Cisco. 2017. White paper: Cisco Vision: 5G-Thriving Indoors. Tech. rep. Cisco. https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/5g-ran-indoor.pdf.

[12] Sébastien Dujardin, Damien Jacques, Jessica Steele, and Catherine Linard. 2020. Mobile phone data for urban climate change adaptation: reviewing applications, opportunities and key challenges. *Sustainability*, 12, 4, 1501.

[13] Joseph C Dunn. 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3, 3, 32–57.

[14] Ericsson. 2017. White paper: Bringing 5G Networks Indoors. Tech. rep. Ericsson. https://www.ericsson.com/en/reports-and-papers/white-papers/bringing-5g-networks-indoors.

[15] European Union. 2016. Eu general data protection regulation (gdpr): regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Retrieved October 18, 2021 from https://gdpr-info.eu/.

[16] Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. 2013. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7, 125–145.

[17] Angelo Furno, Marco Fiore, and Razvan Stanica. 2017. Joint spatial and temporal classification of mobile traffic demands. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, New York City at 3 Park Ave, 1–9.

[18] Angelo Furno, Marco Fiore, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda. 2016. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Trans. Mobile Comput.*, 16, 10, 2682–2696.

[19] Xiaohu Ge, Song Tu, Guoqiang Mao, Cheng-Xiang Wang, and Tao Han. 2016. 5G ultra-dense cellular networks. *IEEE Wireless Commun.*, 23, 1, 72–79.

[20] Mohammadhossein Ghahramani, MengChu Zhou, and Gang Wang. 2020. Urban sensing based on mobile phone data: approaches, applications, and challenges. *IEEE/CAA Journal of Automatica Sinica*, 7, 3, 627–637.

[21] Badih Ghattas, Pierre Michel, and Laurent Boyer. 2017. Clustering nominal data using unsupervised binary decision trees: comparisons with the state of the art methods. *Pattern Recognition*, 67, 177–185.

[22] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: an overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[23] Avi Goldfarb and Jeff Prince. 2008. Internet adoption and usage patterns are different: implications for the digital divide. *Information Economics and Policy*, 20, 1, 2–15.

[24] Eduardo Graells-Garrido, Diego Caro, Omar Miranda, Rossano Schifanella, and Oscar F Peredo. 2018. The WWW (and an H) of mobile application usage in the city: the what, where, when, and how. In *Companion Proceedings of the The Web Conference 2018*. Association for Computing Machinery, New York, NY, USA, 1221–1229.

[25] Kyra H Grantz et al. 2020. The use of mobile phone data to inform analysis of covid-19 pandemic epidemiology. *Nature communications*, 11, 1, 4961.

[26] Yixin Huang, Jiliang Zhang, and Jie Zhang. Wireless channel delay spread performance evaluation of a building layout. arXiv preprint arXiv:2212.05656, (2022).

[27] Huawei. 2018. White paper: Indoor 5G Networks. Tech. rep. Huawei. https://carrier.huawei.com/minisite/Indoor-5G/pdf/Indoor-5G-Networks-White-Paper-V2.0-en.pdf.

[28] Yu Jin, Nick Duffield, Alexandre Gerber, Patrick Haffner, Wen-Ling Hsu, Guy Jacobson, Subhabrata Sen, Shobha Venkataraman, and Zhi-Li Zhang. 2012. Characterizing data usage patterns in a large cellular network. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*. Association for Computing Machinery, New York, NY, USA, 7–12.

[29] K Laursen and C Engedal. 1995. The role of the technology factor in economic growth: a theoretical and empirical inquiry into new approaches to economic growth. *Unpublished MA dissertation*.

[30] Keld Laursen. 2015. Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian business review*, 5, 99–115.

[31] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: a review of machine learning interpretability methods. *Entropy*, 23, 1, 18.

[32] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social media fingerprints of unemployment. *PloS one*, 10, 5, e0128692.

[33] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. Curran Associates Inc., 57 Morehouse LaneRed HookNYUnited States, 4768–4777.

[34] Scott M Lundberg et al. 2020. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2, 1, 56–67.

[35] Bo Ma, Weisi Guo, and Jie Zhang. 2020. A survey of online data-driven proactive 5G network optimisation using machine learning. *IEEE access*, 8, 35606–35637.

[36] Jonas Medbo, Pekka Kyosti, Katsutoshi Kusume, Leszek Raschkowski, Katsuyuki Haneda, Tommi Jamsa, Vuokko Nurmela, Antti Roivainen, and Juha Meinila. 2016. Radio propagation modeling for 5G mobile and wireless communications. *IEEE communications magazine*, 54, 6, 144–151.

[37] Sachit Mishra, Zbigniew Smoreda, and Marco Fiore. 2022. Second-level digital divide: a longitudinal study of mobile traffic consumption imbalance in france. In *Proceedings of the ACM Web Conference 2022* (WWW '22). Association for Computing Machinery, Virtual Event, Lyon, France, 2532–2540. ISBN: 9781450390965. DOI: 10.1145/3485447.3512125.

[38] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. 2020. Explainable k-means and k-medians clustering. In *International conference on machine learning*. PMLR, 7055–7065.

[39] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2, 1, 86–97.

[40] Jorge Navarro-Ortiz, Pablo Romero-Diaz, Sandra Sendra, Pablo Ameigeiras, Juan J Ramos-Munoz, and Juan M Lopez-Soler. 2020. A survey on 5G usage scenarios and traffic models. *IEEE Commun. Surveys Tuts.*, 22, 2, 905–929.

[41] Nuria Oliver et al. 2020. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances*, 6, 23, eabc0764. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abc0764. DOI: 10.1126/sciadv.abc0764.

[42] Marta Orviska and John Hudson. 2009. Dividing or uniting europe? internet usage in the eu. *Information Economics and Policy*, 21, 4, 279–290.

[43] Neeti Pokhriyal and Damien Christophe Jacques. 2017. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, 114, 46, E9783–E9792. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1700319114. DOI: 10.1073/pnas.1700319114.

[44] Julien Ridoux, Antonio Nucci, and Darryl Veitch. 2006. Seeing the difference in ip traffic: wireless versus wireline. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*. IEEE, New York, NY, USA, 1–12.

[45] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.

[46] Ahmed Mohammed Al-Samman, Tharek Abd. Rahman, Tawfik Al-Hadhrami, Abdusalama Daho, MHD Nour Hindia, Marwan Hadri Azmi, Kaharudin Dimyati, and Mamoun Alazab. 2019. Comparative study of indoor propagation model below and above 6 GHz for 5G wireless networks. *Electronics*, 8, 1, 44.

[47] Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, and Thamara Vieira da Rocha. 2021. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. *Transportation Research Part C: Emerging Technologies*, 129, 103183. DOI: https://doi.org/10.1016/j.trc.2021.103183.

[48] Lloyd S Shapley, HW Kuhn, and AW Tucker. 1953. Contributions to the theory of games. *Annals of Mathematics studies*, 28, 2, 307–317.

[49] Jessica E. Steele et al. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14, 127, 20160690. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2016.0690. DOI: 10.1098/rsif.2016.0690.

[50] Iñaki Ucar, Marco Gramaglia, Marco Fiore, Zbigniew Smoreda, and Esteban Moro. 2021. News or social media? socio-economic divide of mobile service consumption. *Journal of The Royal Society Interface*, 18, 185, 20210350. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2021.0350. DOI: 10.1098/rsif.2021.0350.

[51] Marc Verboord. 2017. Internet usage and cosmopolitanism in europe: a multilevel analysis. *Information, Communication & Society*, 20, 3, 460–481.

[52] Xu Wang, Zimu Zhou, Fu Xiao, Kai Xing, Zheng Yang, Yunhao Liu, and Chunyi Peng. 2018. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Trans. Mobile Comput.*, 18, 9, 2190–2202.
[53] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58, 301, 236–244.
[54] Syed Fahad Yunas, Ari Asp, Jarno Niemela, and Mikko Valkama. 2014. Deployment strategies and performance analysis of macrocell and femtocell networks in suburban environment with modern buildings. In *39th Annual IEEE Conference on Local Computer Networks Workshops*. IEEE, New York, NY, USA, 643–651.
[55] Syed Fahad Yunas, Mikko Valkama, and Jarno Niemelä. 2015. Spectral and energy efficiency of ultra-dense networks under different deployment strategies. *IEEE Com. Mag.*, 53, 1, 90–100.
[56] André Felipe Zanella, Orlando E Martínez-Durive, Sachit Mishra, Zbigniew Smoreda, and Marco Fiore. 2022. Impact of later-stages covid-19 response measures on spatiotemporal mobile service usage. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, New York, NY, USA, 970–979.
[57] Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. 2019. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE J. Sel. Areas Commun.*, 37, 6, 1389–1401.
[58] Jiliang Zhang, Andrés Alayón Glazunov, Wenfei Yang, and Jie Zhang. 2021. Fundamental wireless performance of a building. *IEEE Wireless Commun.*, 29, 1, 186–193.
[59] Jiliang Zhang, Andrés Alayón Glazunov, and Jie Zhang. 2021. Wireless performance evaluation of building layouts: closed-form computation of figures of merit. *IEEE Trans. Commun.*, 69, 7, 4890–4906.
[60] Ying Zhang and Ake Årvidsson. 2012. Understanding the characteristics of cellular data traffic. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*. ACM, New York, NY, USA, 13–18.
[61] ZTE. 2019. White paper: 5G Indoor. Tech. rep. ZTE. https://www.zte.com.cn/content/dam/zte-site/res-www-zte-com-cn/mediares/zte/files/newsolution/wireless/ran/white_paper/ZTE_5G_Indoor_White_Paper-EN.pdf.

ETHICS

Our work builds on mobile network traffic generated by users of a nationwide cellular infrastructure, focusing on the traffic generated by ICNs. Specifically, we leverage the hourly traffic demands at the level of individual cells, which are generated from network measurements carried out in the target infrastructure as detailed in Section 3.

The traffic measurements used to compute the measures introduced in Section 4.2 were collected by the operator for network management and research purposes, and temporarily stored within a secure platform at their own premises. The hour-level aggregation was also carried out in the same platform by personnel of the network operator, in full compliance with Article 89 of the General Data Protection Regulation (GDPR) [15] of the European Commission. The data collection and processing were approved by the Data Protection Officer (DPO) of the operator within the context of a collaborative research project.

We remark that the original network measurements contained personal identifiers (*e.g.*, the International Mobile Subscriber Identifier, or IMSI) and sensitive data (*e.g.*, locations of visited cells, or mobile services consumed) about individual users, and were deleted upon aggregation. Instead, the aggregated traffic demands used (over the two-month period or the hourly patterns presented in Section 6) do not contain personal identifiers or sensitive information. The level of spatio-temporal aggregation ensures that no data subject can be re-identified, and that the traffic data or the metric used for clustering do not configure as personal data in the GDPR acceptation.

The researchers involved in the work presented in this paper only had access to such aggregated and privacy-preserving traffic data for the purpose of carrying out the study. Ultimately, our dataset and research do not involve risks for the mobile subscribers, while they provide new knowledge about the dynamics of indoor mobile traffic demands, which will benefit an improved design and more dependable validation of technical solutions for mobile network operations.