An Experimental Study of Factor Analysis over Cellular Network Data

Feyzullah Kalyoncu, Engin Zeydan, Ahmet Yildirim and Ibrahim Onuralp Yigit
Türk Telekom Labs, Istanbul, Turkey 34889
E-mail: {feyzullah.kalyoncu, engin.zeydan, ibrahimonuralp.yigit}@turktelekom.com.tr, ahmet.yildirim3@boun.edu.tr

Abstract—Mobile Network Operators (MNOs) are evolving towards becoming data-driven, while delivering capacity to collect and analyze data. This can help in enhancing user experiences while empowering the operation workforce and building new business models. Mobile traffic demands of users can give insights to MNOs to plan, decide and act depending on network conditions. In this paper, we investigate the behaviour of Istanbul residents using the cellular network traffic activity over spatial and temporal dimensions via exploratory factor analysis (EFA) using a major MNO's cellular network traffic data in Turkey. Our results reveal various time and spatial patterns for Istanbul residents such as morning and evening commuting factors, business and residential factors as well as nightlife and weekend afternoon factors as the most prominent cultural behaviour. The analysis results also demonstrate interesting findings such as tunnels and transportation paths selected by Istanbul residents may differ during morning rush work hour compared to evening rush after-work hour.

Index Terms—factor analysis, cellular data, spatio-temporal, mobile operators

I. INTRODUCTION

In the era of digital age, data can inform and empower transformations of all industry such as telecommunications, health, automotive and factories of future. Using the data as a commodity, organization institutions can detect important phenomena as early as possible, can forecast the outcomes that are yet to come (predictive) or can perform optimization for better modeling and outcomes (prescribe). Some of the enablers for this kind of process enhancement and gaining enterprise level intelligence from rich dataset are via machine learning and statistical methodologies [1]. The cellular data collected from various locations at different time provides rich and important context data to Mobile Network Operators (MNOs) for better decision making and operational process enhancements. For example, the collected data can provide insights into different mobility patterns for groups of subscribers over the observed duration periods after data analysis. This can enhance the capabilities of MNOs for providing personalized services for different customer segments based on their personalized experiences. The mobile data interaction can reflect the personal behaviour, preferences and objectives of users. For this reason, MNOs are working to improve operational efficiency, obtain predictive analytics and extract insights by analyzing massive dataset available in their premises.

For controlling and optimization of network operations, historical and contextual data analysis for modeling the traffic at cell level such as [2], [3], [4], [5] and profiling the user behaviour at different time scales such as [6], [7] based on mobile traffic demand exist in the literature. An analysis that relies on Exploratory Factor Analysis (EFA) using with real-world mobile traffic dataset for Milan and Paris cities are performed in [6]. The results reveal different network activity profiles in those two major European cities. In this paper, we further extend the EFA presented for Milan and Paris cities and study network activity profiles of Istanbul using real-world mobile traffic dataset of a major MNO in Turkey. Compared to previous works in mobile data analysis, our results reveal various mobile usage demands of Istanbul residents and also shed some light on cultural behaviour.

II. SYSTEM MODEL AND ARCHITECTURE

The proposed platform processes and analyzes network key performance indicator (KPI) data that is collected a priori from Operations Support System (OSS) of the MNO in Turkey. Fig. 1 demonstrates the general architecture of our utilized solution. The proposed architecture is composed of three main modules: (a) Data Collector Tier, (b) Analysis Tier, (c) Visualization Tier. The solution is based on utilization of the network KPI data together with open-source data analytics software and platforms. Pandas and R packages are used as data analysis tools utilizing data-centric packages. For map visualization, Folium visualization tool [8] and inside the Folium, Leaflet javascript [9] library are utilized.

![Fig. 1: System Architecture for EFA and data visualization](image)

In Data Collector Tier marked as step-(1) in Fig. 1 the data is collected from OSS similar to Fig. 2 and is transferred...
into Pandas and R data analytics tools marked as step-(2). In Pandas, first the data in csv format is grouped by Site-ID and the latitude, longitude of the sites are appended into the existing csv data format. In step-(2), all the packet switched (PS) traffic corresponding to each Cell-ID for 4G are analyzed. An example of sum 4G traffic values for Besiktas and Umraniye District of Istanbul is given in Table I. Finally, a matrix with each Cell-ID as row and hourly median PS traffic values over a week in 4G as column matrix is constructed and fed into EFA for determining different factors from the cellular data traffic. After EFA, different factors and corresponding Base Station (BS) scores for each factor are obtained. Later, the analysis results are visualized in Visualization Tier, marked as step-(3) in Fig. 1 where we utilize Folium & Leaflet maps in order to visualize the EFA scores of each cell sites for each obtained factor.

**Model:** Let $\mathbf{X}$ be a $N \times 1$ vector of observed variables, i.e. phenomena of interest such as Download (DL), Upload (UL) traffic or number of users on a given BS in this paper. The fundamental equation for factor analysis is defined as [6]

$$\mathbf{X} = \mathbf{AF} + \mathbf{U}$$

(1)

where $\mathbf{A}$ represents $N \times K$ matrix of common factor pattern coefficients that describe importance of each factor to every variable, $\mathbf{F}$ represents $K \times 1$ vector of unknown normalized common factors, i.e., a small number ($K<<N$) of complex relationships among variables and $\mathbf{U}$ represents $N \times 1$ vector of unknown unique factors that are specific to a single variable. Hence, $\mathbf{A}$ is a weighted combinations of the common factors in $\mathbf{F}$ and the unique factors in $\mathbf{U}$. Together with EFA solution, by analyzing variable observations from a set of samples, EFA can identify common/unique factors, and numerical relationships that describe how much each common factor explains each variable [10].

In our analysis, in order to determine the number of factors to extract, we have utilized *parallel analysis* where the largest eigenvalues of the data correlation matrix is selected. For factor rotation, we have selected *promax* for maximizing the high loadings using R project [11].

TABLE I: An example of sum 4G traffic values for Besiktas and Umraniye District of Istanbul.

| District           | DL Traffic (GB) | UL Traffic (GB) | No. of Users (avg.) |
|-------------------|-----------------|-----------------|---------------------|
| Besiktas          | 4,239,906.386   | 488,613.8446    | 67,460.33           |
| Umraniye          | 9,460,358.224   | 870,681.5689    | 142,626.2021        |

TABLE II: Statistics of Analyzed Cellular Data in Istanbul.

|                   |                  |
|-------------------|------------------|
| # of rows         | 6,264,286        |
| # of districts    | 40               |
| PS DL traffic     | 18,318.87544     |
| PS UL traffic     | 1,793.451409     |
| Obs. Duration     | 1 month          |
| Average # of users| 284.5            |
| N, K              | $7 \times 24$, 10-13 |
TABLE III: Factors Analysis Descriptions in Istanbul

| Factor  | Labeled Areas                              | Description                                                                 |
|---------|--------------------------------------------|-----------------------------------------------------------------------------|
| DL 1    | Residential Areas                          | Large population residential areas in western part of European side and eastern part of Anatolian side |
| No. of Users 1 | Office, Campus and Industrial Areas | University campuses and business zones in Maslak, industrial areas around Basaksehir and commercial areas around Besiktas, Fatih in European side and Uskudar and Kadikoy (commercial regions) and Dudullu (industrial regions) in Anatolian side |
| No. of Users 2 | Malls, Touristic Areas and Leisure Activity | Historical Regions (Sultanahmet, Kapalicarsi, Galata tower), Shopping centers in Maslak and Bakirkoy, Leisure time activity on coastal side and Bosphorus view sites |
| DL 2    | Morning Commuting                          | Commuting traffic from Anatolian side into European side over bridges, Metrobus line in European side and Avrasya tunnel entrance and exit points |
| DL 3    | Evening Commuting                          | Metrobus express bus line in European side and Highway hub in Anatolian side |
| DL 4    | Farmer’s market, Major Bus Terminal and Airport, night life area | Wholesale food market and Ataturk airport and nightlife (Taksim, Besiktas area) |

We can observe that e.g. for Number of User Factor 1 of Fig. 3c there exists high traffic between 9 am to 6 pm from Monday to Friday, whereas for DL Factor 3 of Fig. 4c, high network utilization exists between 6 am to 9 am. Table III explains the observed factor analysis labels and corresponding location descriptions in Istanbul. From Fig. 3 we can observe how the cells belonging to DL Factor-1 of Fig. 3a and Fig. 3b are highlighting the areas that are mostly populated with local residents, where most of the residential areas are located. Most of the network activity is around residential centers such as Bagcilar, Gungoren, Alibeykoy in European side and Umraniye, Sultanbeyli in Anatolian side. Geographical areas where Number of Users Factor 1 as shown in Fig. 3c and Fig. 3d are mostly related to university campuses, office and industrial areas of Istanbul. This is in line with the properties of Number of Users Factor 1 of Fig. 3c which characterize the working hour traffic between 8 am to 5 pm belonging to BSs whose mobile data traffic activity surges during working hours. Number of Users Factor 2 of Fig. 3e and Fig. 3f on the other hand, characterize the touristic, shopping and leisure activity areas in Istanbul. The historic city center Sultanahmet, Kapalicarsi, Galata Tower that are major touristic attractions for foreigners and local tourists, big shopping malls for local residents such as Istinye Park in Maslak (northern European side), Cevahir and Zorlu Center in Levent (middle European side), Mall of Istanbul, Ataturk airport shopping center (AVM), Marmara Forum, Galeria, Forum Istanbul, Ike, Atakoy Plus (southern European side) are highlighted in Number of Users Factor 2 of Fig. 3e and Fig. 3f. Another dimension in Number of Users Factor 2 that overlaps with the shopping mall activities is indication of the leisure occupations of Istanbul residents around both Anatolian and European coastal sides as well as Bosphorus view regions.

We can also observe different patterns from the other factor profiles. DL Factor 2 and DL Factor 3 of Fig. 3 show the commuting behaviour of Istanbul residents. In Istanbul, most of the residents live in Anatolian side and commute to European side of Istanbul for work. Therefore, after work between 5-8 pm of Fig. 4a and Fig. 4b there exists coherent usage of public transport services, e.g. Metrobus express bus line used by local commuters is running in major locations of Istanbul can be clearly observed from Fig. 4b. During morning commuting hours between 7 am to 9 am of Fig. 4c and Fig. 4d the scores are somehow distributed around entrance from Anatolia side into European side due to existence of non-uniform starting hours of businesses (ranging from 6 am to 9 am) and Metrobus line is less visible compared to evening commuting. In DL Factor 3 of Fig. 4c and Fig. 4d other active areas that morning commuters frequently use are entrance and exit points of Avrasya tunnel (which runs under Marmara sea between Anatolian side and European side) where morning commuters enter from Anatolian side and exist into European side. We can also notice the interesting fact that commuter behaviour around usage of Avrasya tunnel is only visible in morning commuter behaviour, not in evening commuter behaviour.

The areas highlighted by DL Factor 4 of Fig. 4c and Fig. 4f show the activities between 2-4 am where major bus station, major airport, wholesale market hall as well as nightlife areas in Istanbul are highlighted around those times. The geographical cells that have high scores belonging to DL Factor 4 demonstrate wholesale market hall where all the goods (fruits and vegetables) are stocked for servicing the next day. DL Factor 4 is also indicating nightlife occupations of Istanbul residents that overlap with terminal (bus and airport) activities.

IV. CONCLUSIONS

In this paper, we have investigated hourly cellular traffic data and performed factor analysis over the collected dataset of one month of a major cellular mobile network operator in Turkey. The results reveal that there exists different patterns of traffic over different factors depending on the day of the week as well as time of the day. Our results have revealed major residential, business, touristic and market areas as well as morning and evening commuter paths of residents in Istanbul depending on the behaviour of cellular network usage over
Fig. 3: (a) DL Factor 1 (b) DL Factor 1 Map (c) Number of Users Factor 1 (d) Number of Users Factor 1 Map (e) Number of Users Factor 2 (f) Number of Users Factor 2 Map

different time zones. As a future work, we are working on extending the analysis over one year period that can also cover the special events and holidays.

REFERENCES
[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, “Large-scale mobile traffic analysis: a survey,” IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 124–161, 2016.
[2] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in INFOCOM, 2011 Proceedings IEEE, pp. 882–890, IEEE, 2011.
[3] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, “Approaching the limit of predictability in human mobility,” Scientific reports, vol. 3, 2013.
[4] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of predictability in human mobility,” Science, vol. 327, no. 5968, pp. 1018–1021, 2010.
[5] O. Narmanlioglu, E. Zeydan, M. Kandemir, and T. Kranda, “Prediction
Fig. 4: (a) DL Factor 2 (b) DL Factor 2 Map (c) DL Factor 3 (d) DL Factor 3 Map (f) DL Factor 4 (g) DL Factor 4 Map

of active ue number with bayesian neural networks for self-organizing lte networks,” in Network of Future (NoF), 2018 Proceedings, London, UK, 2017.

[6] A. Furno, M. Fiore, and R. Stanica, “Joint spatial and temporal classification of mobile traffic demands,” in INFOCOM–36th Annual IEEE International Conference on Computer Communications, 2017.

[7] D. Naboulsi, R. Stanica, and M. Fiore, “Classifying call profiles in large-scale mobile traffic datasets,” in INFOCOM, 2014 Proceedings IEEE, pp. 1806–1814, IEEE, 2014.

[8] “Folium: Python Data Visualization.” [https://folium.readthedocs.io/en/latest/](https://folium.readthedocs.io/en/latest/) 2017. [Online; accessed 06-Dec.-2017].

[9] “Leaflet: an open-source JavaScript library.” [http://leafletjs.com/](http://leafletjs.com/) 2017. [Online; accessed 06-Dec.-2017].

[10] S. A. Mulaik, Foundations of factor analysis. CRC press, 2009.

[11] “Principal Components and Factor Analysis.” [https://www.statmethods.net/advstats/factor.html](https://www.statmethods.net/advstats/factor.html) 2018. [Online; accessed 19-Jan.-2018].