Traffic Overload Estimation in Telecommunications Network Using Trajectory Data Analysis

Mohammad Hossein Zarei [1*], Ali Zare Zardiny [2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran - m.hosseinzarei@ut.ac.ir
[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran - zare_zardiny@ut.ac.ir

Keywords: Spatiotemporal Data, Telecommunications Network, Network Congestion, Trajectory Data, Random Forest Regressor

Abstract

In modern telecommunication networks, traffic congestion and overload on cells pose a serious challenge to maintaining service quality. While nominal capacity values for each cell are often known, operators typically lack accurate insights into how much traffic exceeds these thresholds in congested areas. In this study, an approach is proposed to estimate the excess traffic load on each cell using trajectory data. The core idea is that dynamic population density, derived from the spatial distribution of users, plays a crucial role in shaping traffic patterns. To approximate each cell's coverage area, circular approximation is employed, allowing users to be associated with their nearest cell based on spatial proximity. By combining user presence per cell and hour with simulated traffic values, a comprehensive dataset is constructed and fed into a Random Forest Regressor model. The trained model is used to predict actual traffic, and overload conditions are detected by comparing predicted values with nominal capacities. The proposed framework enables identification of congestion-prone cells and supports informed decision-making for network infrastructure expansion.

1. Introduction

In recent years, the widespread use of mobile devices and the internet has led to a significant increase in telecommunications network traffic. One of the most critical challenges in this field is the identification and estimation of traffic exceeding the capacity of Base Transceiver Station (BTS) antennas, a condition in which actual traffic surpasses the nominal capacity of the antenna, resulting in service quality degradation, data transmission delays, and user dissatisfaction.

In such cases, not only is the excess traffic not clearly defined, but the additional load that the network must handle also remains uncertain. Similar to any business or commercial environment, estimating the demand for deliverable services in the telecommunications sector can greatly contribute to effective network planning and optimization (Shiomoto, 2023). One of the key factors directly influencing network traffic growth is the presence of users in various geographic areas.

Although demographic data is typically provided in the form of static population figures within urban blocks, in reality, dynamic population plays a crucial role in network traffic analysis. Trajectory data, as a novel and valuable resource, enables the analysis of dynamic population. This data includes spatiotemporal information about individuals' movement paths, allowing for accurate identification of user presence patterns and density around BTS antennas.

The main objective of this study is to estimate and predict excess traffic beyond the capacity of telecommunication antennas based on trajectory data. In this research, using trajectory data, a model is developed for estimating telecommunications network traffic. The remainder of this paper is organized into four sections aligned with the defined objective. Section 2 reviews previous related studies. Section 3 outlines the theoretical foundations. Section 4 introduces the research methodology. Section 5 presents the model implementation process and its resulting outcomes.

Finally, analysis of the results, the overall conclusion, and recommendations for future research are discussed.

2. Theoretical Foundations

Given that the focus of this study is on traffic management and modeling within a 4G network, it is essential first to identify the key performance indicators (KPIs) relevant to this network. These KPIs are as follows:

1. **Initial E-RAB Establishment Success Rate**
   This KPI represents the success rate of the initial E-RAB (Evolved Radio Access Bearer) establishment process. This KPI is categorized under accessibility metrics (Ferreira, 2020). A decrease in this indicator may result from network congestion caused by an increased number of users.
2. **Total Traffic**
   This parameter is calculated by measuring throughput over time and refers to the total volume of data transmitted across the network, including both downlink and uplink data, within a specific time frame (e.g., one hour or one day). Equation 1 illustrates the method used to calculate this parameter (3GPP TS 32.425 version 10.6.0 Release 10, 2012).

$$Total\ Traffic = \int_{t_1}^{t_2} PDCP\,UL\ Throughput(t) + PDCP\,DL\ Throughput(t)\ dt \quad (1)$$

where t_1 = start time of the measurement period
t_2 = end time of the measurement period

[*] Corresponding author

PDCP UL Throughput = uplink throughput at the PDCP layer
PDCP DL Throughput = downlink throughput at the PDCP layer

3. **RRC Connection Setup Success Ratio**
   This KPI measures the success rate of the RRC (Radio Resource Control) connection setup process, which is the stage where the User Equipment (UE) initiates a connection request to the eNodeB (evolved NodeB) via an RRC Connection Request message (Hendrawan, 2019). This KPI is also considered one of the most critical indicators for evaluating network accessibility (Ferreira, 2020). Similar to the first KPI, this metric reflects a consequence of reduced network accessibility (rather than its cause) and is influenced by the number of users and the utilization of network resources. In fact, a decrease in a cell's accessibility will also lead to a drop in this KPI.
4. **Call Drop Rate**
   This indicator generally refers to the phenomenon of call or data packet disconnection in voice and data networks. A call drop or packet loss occurs when an ongoing session is unexpectedly terminated before either party intentionally ends it. Mobile service operators use this KPI to assess the quality of service (QoS) (Tarkaa, 2011).
5. **PDCP Layer Throughput DL**
   This KPI measures the downlink data transmission rate in the PDCP (Packet Data Convergence Protocol) layer for a specific cell (3GPP TS 32.425 version 10.6.0 Release 10, 2012). It reflects the average throughput of PDCP SDUs in the download direction, measured in megabits per second (Mbps), and is used to evaluate the network's quality and efficiency in delivering data to the UE.
6. **PDCP Layer Throughput UL**
   This KPI measures the uplink data transmission rate in the PDCP layer for a specific cell (3GPP TS 32.425 version 10.6.0 Release 10, 2012). It reflects the average throughput of PDCP SDUs (Service Data Units) in the upload direction, measured in Mbps, and is used to assess the network's quality and efficiency in transferring data from the UE to the eNodeB.
7. **User Throughput DL**
   This indicator is considered the most representative KPI for assessing the user's experience (Mostafa, 2022). It shows the actual data rate (in Mbps) at which the user receives data from the eNodeB to their device.
8. **User Throughput UL**
   This metric represents the actual data rate (in Mbps) at which the user sends data from their device to the eNodeB.

Given the diversity and breadth of KPIs and network quality assessment indicators, Total Traffic stands out as a more comprehensive metric compared to others and is therefore the focus of this study.

In telecommunication networks, each antenna has a designated nominal capacity, defined based on factors such as network design, signal power, and available bandwidth. This is referred to as Nominal Traffic. However, under real-world conditions, especially during peak hours, network usage can exceed this nominal capacity. The actual amount of traffic consumed by users at any given time is called Actual Traffic. When actual traffic surpasses the nominal threshold, it results in Overloaded Traffic, which may lead to degraded service quality, increased latency, or even connection failures.

In such circumstances, the traffic exceeding capacity triggers network congestion. Congestion occurs when a network link or node carries more data than it can handle, such that the available path capacity is insufficient for smooth data transmission. This condition leads to reduced data flow, queuing delays, packet loss, and ultimately, a decline in the quality of service delivery (Dike, 2013).

Despite these challenges, network operators typically lack the means to accurately quantify the excess traffic that exceeds an antenna's capacity. This inability to assess the precise extent of overload hampers informed decision-making regarding infrastructure expansion or capacity upgrades.

Without a quantitative understanding of the excess demand, it becomes difficult to prioritize the installation of new antennas or the enhancement of existing resources effectively.

Since network traffic levels in a given area are influenced by the population of users present, analyzing the spatiotemporal distribution of users within the coverage area of each antenna can provide valuable insights into network load. Accordingly, this study introduces the idea of leveraging trajectory data as a potential solution.

Trajectory data is a form of spatiotemporal information that records users' movement paths over time, typically including user identifiers, timestamps, and spatial coordinates. Such data can serve as a complementary source, offering a detailed view of dynamic population density and its distribution across time and space. By estimating the real-time population surrounding each antenna, this data enables more accurate modeling of the antenna's traffic load.

3. Literature Review

Spatial data is among the most influential data types in decision-making and optimization processes within telecommunication networks, and has therefore been the focus of numerous studies. This section reviews research efforts that have examined the use of spatial data for optimizing telecommunication networks.

Amiri (2023) investigated the optimization of a location-allocation model for deploying antennas in a telecommunication network in the northern region of Kermanshah. The study aimed to propose an effective solution for identifying optimal antenna locations using a Genetic Algorithm (GA). To achieve this, a GA-based model was proposed to improve the spatial coordinates of existing antennas extracted from a Geographic Information System (GIS).

Wang (2020) examined the optimization of antenna locations in ultra-dense 5G networks. Several blocks in Wuhan, China, were selected as the study area. The data sources included road and pathway data from OSM, building data from AliMap, and 4G site coordinates obtained from a telecommunications operator.

Adegboyega (2017) assessed the suitability and vulnerability of telecommunications antennas based on GIS in the city of Ibadan, Nigeria. This research utilized GIS techniques to evaluate the spatial distribution of antennas, identify vulnerable zones in terms of health hazards, and determine suitable locations for new installations.

Tayal (2017) investigated suitability analysis for determining the optimal location of telecommunication antennas using GIS in Uttarakhand, India. The findings indicated that integrating spatial layers with population data as a key parameter could justify the need for new antennas. Consequently, this approach supports proposing new antenna locations and performing spatial overlap analysis.

Although various studies have addressed the use of a wide range of spatial data, the application of user trajectory data for optimizing telecommunication networks remains a relatively underexplored area, an aspect that this study aims to highlight.

4. Methodology

In this study, the goal is to estimate the actual traffic load on each cell using users' trajectory data. This estimation serves as a basis for analyzing the amount of overload relative to the nominal capacity of each cell and ultimately identifying cells experiencing congestion in the network. Figure 1 display an overview of the proposed methodology.

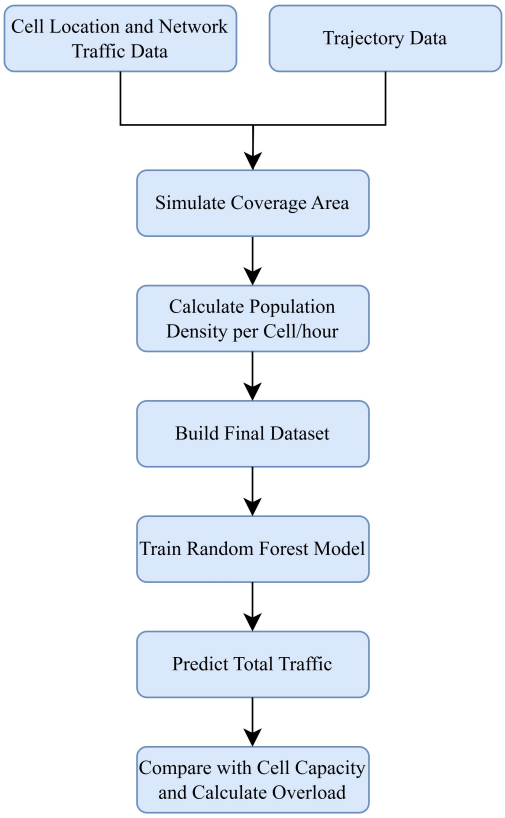

Figure 1. Flowchart of the research procedure

In the first step, the spatial location data of cells along with their traffic data and users' trajectory data are collected.

In the second step, the coverage area of each cell is simulated. This coverage area can be modeled using various methods such as Voronoi diagram (NCHRP research report 868, 2018; Fekih, 2019) or circular approximation. Then, by defining the service areas and user locations, the process of mapping users to their nearest antenna is performed.

In the third step, the user density within each cell coverage area is calculated for each time interval.

In the fourth step, data including cell ID, spatial location of each cell, time, user density within each cell, and total traffic of each cell are integrated into a single dataset. This dataset is used as input to the machine learning model.

In the fifth step, modeling is conducted. This paper employs the Random Forest Regressor algorithm, which effectively captures complex nonlinear dependencies between population density, spatial variables, and temporal variations, inherent characteristics of telecommunication traffic patterns. Its ensemble structure, based on multiple decision trees, enhances prediction accuracy and robustness while minimizing the risk of overfitting. In addition, Random Forest provides interpretability by quantifying the relative importance of each feature, helping to explain how factors such as user density and time contribute to variations in network load. These capabilities make it particularly suitable for predicting network behavior and assessing the influence of dynamic population on traffic.

Using the Random Forest Regressor model, the relationship between influential features such as population density, spatial location, and time with the total traffic value is learned. In the training process, features and target variables are defined.

To evaluate model performance, the available data are split into training and testing subsets. The machine learning model is then trained with the training data to estimate the actual traffic of each cell at different times based on input variables.

In the sixth step, the trained model is applied to predict the traffic for each cell over various time intervals. This prediction represents the approximate traffic volume entering each cell under current conditions.

In the final step, the predicted traffic is compared with the nominal capacity of each cell. If the traffic exceeds this capacity, the overload amount is calculated, and cells experiencing congestion are identified.

5. Implementation

This section presents the technical details and the complete process of implementing the proposed solution. The data used in this implementation are simulated datasets employed in a test project. Parts of the datasets, shown in Figure 2.

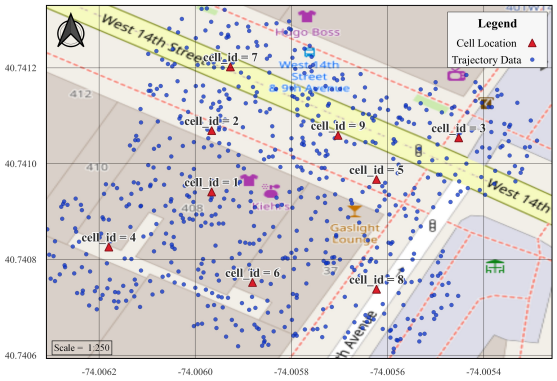

Figure 2. Simulated datasets

Two simulated datasets, trajectory and traffic, were loaded in Python. After converting the time columns to the standard datetime format, the hour component was extracted from these columns. To calculate the user density within the coverage area of each cell, first, using circular approximation, each user's location was compared with a defined radius around the cells.

Then, for each hour, the number of users present near each cell was counted and added to the cell data as the variable population_density. The output of this step is a final dataset containing spatial, temporal, and population information for each cell.

After data preparation, input features and output target variables for model training were determined. The input features included population density in each cell, hour of the day, and geospatial coordinates of the cell (longitude and latitude).

These features were selected based on their direct and logical relationship with telecommunication traffic load. Dynamic

population density represents the most influential factor, as an increase in the number of users within the service area of each antenna directly leads to a higher volume of exchanged data and, consequently, increased traffic. The time feature (hour) reflects users' temporal behavior patterns, since network traffic typically fluctuates throughout the day; for example, it tends to be higher during working hours compared to late-night periods. Combining this feature with population density enables the model to capture temporal variations in user demand more effectively. Additionally, the geospatial coordinates are included to account for spatial differences among cells. These variables allow the model to identify subtle spatial variations in antenna traffic behavior, even between neighboring coverage areas.

The target variable was total traffic, i.e., the actual recorded traffic for each cell at each hour. For training and evaluating the model, the data were split into training and testing sets, with 70% of the data allocated for training and 30% for testing.

The model used was the Random Forest Regressor algorithm from the scikit-learn library. To optimize the performance of the model and prevent overfitting, the hyperparameters were tuned in a stepwise manner. Initially, the baseline model was trained with 100 decision trees (n_estimators=100) using a fixed random seed (random_state=42) to ensure stability and reproducibility of the results. After evaluating the initial results, the number of estimators was increased to 200 to enhance prediction stability and reduce random fluctuations. Additionally, the maximum tree depth (max_depth) was limited to 20 to control model complexity and prevent overfitting. These parameters were selected after testing several combinations to achieve a balance between model accuracy and generalization capability. The final configuration provided more stable and accurate predictions compared to the initial setup.

Based on the trained model, the traffic exceeding the nominal capacity of each antenna can be determined by comparing the predicted total traffic with the nominal capacity, enabling identification of critical points in the network.

For example, if the nominal capacity of each cell is considered as 250 Mbps, the excess traffic for each cell and its occurrence time can be predicted. The resulting outcomes are presented in Table 1.

Cell_id	Hour	Predicted Traffic	Nominal Capacity	Overload
1	14	275.92	250	25.92
4	12	296.66	250	46.66
1	13	321.11	250	71.11
2	12	279.19	250	29.19
9	12	265.03	250	15.03
5	17	299.56	250	49.56
2	14	294.74	250	44.74
2	23	257.5	250	7.5
5	14	277.88	250	27.88
6	14	263.76	250	13.76
3	12	281.58	250	31.58
4	13	273.14	250	23.14
7	14	282.2	250	32.2
9	15	306.95	250	56.95

Table 1. Overload traffic estimation per cell and hour

Figure 3 illustrates the temporal and spatial distribution of overload traffic in the network cells in a two-dimensional format. In this chart, the horizontal axis represents the hours of the day during which overload occurred, and the vertical axis indicates the cell IDs within the network.

The color intensity in each cell of the matrix corresponds to the amount of overload traffic (in Mbps) for that specific cell and hour, with darker colors indicating higher levels of overload. By utilizing this heatmap, spatiotemporal patterns of congestion hotspots can be easily identified.

For instance, if several cells repeatedly experience overload during midday hours (e.g., between 12:00 and 17:00), this may indicate a geographic concentration of users or high data demand during business hours. Conversely, cells exhibiting overload during low-traffic hours (such as midnight) may be located in densely populated areas or zones with nighttime activity.

Such insights can serve as a basis for decision-making regarding optimization of radio resource allocation, capacity enhancement, or deployment of new antennas in critical locations.

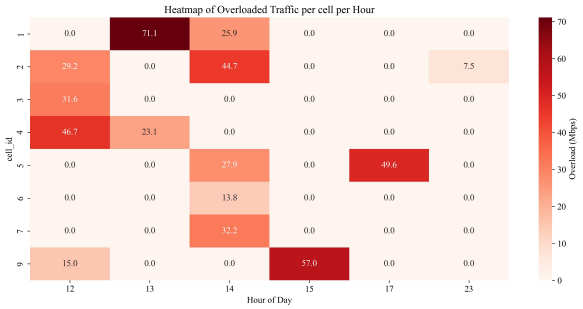

Figure 3. Heatmap of overloaded traffic per cell and hour

It should be noted that all implementation steps were carried out using Python version 3.12. For data processing and model implementation, in addition to the scikit-learn library, pandas, numpy, geopy, matplotlib, seaborn, and joblib libraries were utilized.

6. Interpretation of Results

After training the model, predictions were performed on the test dataset, and the model's performance was evaluated using common metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. The evaluation results, shown in Table 2.

Evaluation metric	Value
MAE	36.28 Mbps
MSE	1800.75 Mbps
RMSE	42.44 Mbps
R^2 Score	0.78

Table 2. Evaluation metrics

Comparison of MAE and RMSE indicates that the model has an average prediction error of approximately 40 Mbps, while larger errors measured by RMSE are around 42 Mbps. The coefficient of determination ($R^2 = 0.78$) signifies that the model can explain about 78% of the variance in actual traffic using the specified features. These values demonstrate the model's satisfactory performance with acceptable accuracy within the simulated range.

One advantage of the Random Forest Regressor is its ability to assess the relative importance of features. After training, the

importance of each of the four features was examined. The results, shown in Table 3.

Feature	Importance
Population Density	0.529
Hour	0.391
Latitude	0.045
Longitude	0.033

Table 3. Feature importance in the random forest regressor

Result shows that population density has the highest influence on traffic prediction, contributing about 53%. This finding aligns with the theoretical assumption of the research, as an increase in the number of users near an antenna is the primary factor increasing the load and traffic generated by that antenna. Following this, the time of user presence (hour) holds nearly 40% importance, with geospatial coordinates of antennas ranked thereafter.

The relatively low importance of geospatial coordinates in the model does not imply that the antennas' locations are irrelevant; rather, it reflects the nature of the data and the degree of spatial homogeneity within the study area. Since all examined cells are situated within a relatively small portion of an urban area, the spatial differences between antennas in terms of geographic coordinates are minimal, and traffic variations are predominantly influenced by dynamic population density and users' temporal patterns rather than position.

In other words, at such a limited spatial scale, the antennas' coordinates do not serve as a strong differentiating factor for traffic prediction, as all network points are embedded in an environment with similar urban morphology and demand characteristics. If the study were conducted at a larger scale (e.g., across multiple cities or regions with diverse characteristics), the importance of geographic location in the model would be expected to increase.

The significant importance of user presence time indicates that the temporal distribution of user movement meaningfully affects traffic distribution.

For a deeper understanding of model performance, the analysis of the two charts presented in Figures 4 and 5 can be considered. Figure 4 shows a scatter plot of actual values (horizontal axis) versus predicted values (vertical axis) for test samples. The red diagonal line defined by y = x represents perfect predictions where all points lie exactly on the line. It was observed that most points cluster around this line with a relatively symmetrical distribution of scatter; this implies that the model tends to produce predictions close to the true values and there is no significant deviation over most points.

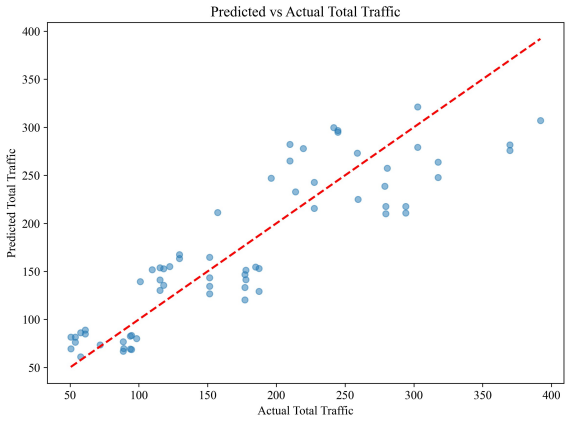

Figure 4. Scatter plot of predicted vs. actual traffic values

For a more detailed analysis, a subset of the test data was plotted as two separate lines representing the actual and predicted values. Figure 5 demonstrates that the traffic variations across the sample indices are generally synchronized, with discrepancies observable only at a few points. In most cases, the actual and predicted lines are closely aligned, indicating the model's capability to accurately track traffic fluctuations.

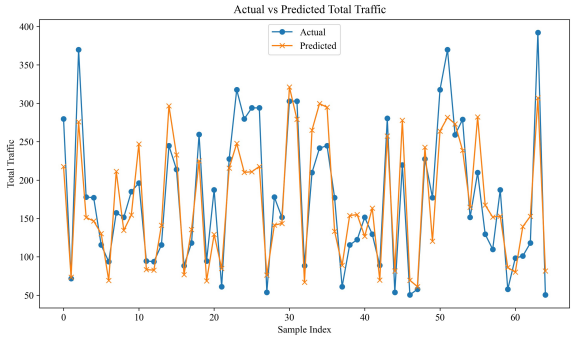

Figure 5. Comparison of predicted and actual traffic over sample index

Based on the quantitative results and qualitative plots, it can be concluded that the Random Forest Regressor model has effectively captured the correlation between dynamic population density and network traffic volume, while accounting for temporal and spatial trends in traffic prediction.

7. Conclusion

This study proposed the use of trajectory data for modelling network traffic load. In the implementation phase, a model based on the Random Forest Regressor algorithm was developed to address a key challenge in telecommunication network optimization: estimating traffic load exceeding the nominal capacity of cells. This modelling combined features such as dynamic population density, time, and location to predict the traffic load of each cell.

The results and evaluations indicate that integrating users' spatial data with traffic prediction models offers an effective approach for identifying high-traffic areas and enabling intelligent decision-making in telecommunications network development.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-849-2026 | © Author(s) 2026. CC BY 4.0 License.

However, this study faced several limitations that warrant further investigation in future research. The dataset used was simulated; thus, utilizing real operator data could enhance model validity and improve the reliability of the outcomes. Additionally, this research employed a circular approximation to estimate cell coverage areas, which does not accurately represent the network behaviour in serving cell selection. Access to actual Serving Cell data would allow for a more precise allocation of traffic to cells. Furthermore, the nominal cell capacities were considered fixed and predefined, whereas in practice, capacity may vary depending on factors such as allocated bandwidth, technology, and instantaneous network load.

References

3GPP TS 32.425 version 10.6.0 Release 10, 2012: Telecommunication management; Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN).

Adegboyega, S.A.A., et al, 2017: GIS-based site suitability and vulnerability assessment of telecommunication base transceiver station facilities in Ibadan metropolis, Nigeria. *journal of the Italian Society of Photogrammetry and Topography (SIFET)*. doi.org/10.1007/s12518-017-0194-y.

Amiri, F., 2023: Optimization of facility location-allocation model for base transceiver station antenna establishment based on genetic algorithm considering network effectiveness criteria (case study north of Kermanshah). *journal of Scientia Iranica*. doi.org/10.24200/sci.2021.55207.4116.

Dike, D. O., Iroh, C. U., Ezeh, G. N., Dike, B. C., Chukwuchekwa, N. and Ndinechi, M. C., 2013: Minimization of call congestion in telecommunication system using OFDM optimization model. *IEEE International Conference on Emerging & Sustainable Technologies for Power & ICT in a Developing Society (NIGERCON)*, Owerri, Nigeria, 2013, pp. 123-128. doi.org/10.1109/NIGERCON.2013.6715646.

Fekih, M., Bonnel, P., Smoreda, Zb., Bellemans, T., Furno, A., Galland, St., 2019: Méthodologie de filtrage et de traitement de données de signalisation de la téléphonie mobile pour la construction de matrice origine-destination. Application à la région rhône-alpes. *Les Cahiers Scientifiques du Transport* / Scientific Papers in Transportation, 2019, 75, pp.81-111. hal-04185213.

Ferreira, D., et al, 2020: Root Cause Analysis of Reduced Accessibility in 4G Networks. pages 129-145. In: Boumerdassi, S., Renault, É., Mühlethaler, P. (eds) Machine Learning for Networking. MLN 2019. *Lecture Notes in Computer Science(), vol 12081. Springer, Cham*. doi.org/10.1007/978-3-030-45778-5_9

Hendrawan, H., 2019: Accessibility Degradation Prediction on LTE/SAE Network Using Discrete Time Markov Chain (DTMC) Model. *Journal of ICT Research and Applications*. 13. 1-18. 10.5614/itbj.ict.res.appl.2019.13.1.1.

Mostafa, A., Elattar, M. A., Ismail, T., 2022: Downlink Throughput Prediction in LTE Cellular Networks Using Time Series Forecasting. *International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom)*, Graz, Austria, 2022, pp. 1-4. doi.org/10.1109/CoBCom55489.2022.9880654.

National Academies of Sciences, Engineering, and Medicine. National Cooperative Highway Research Program, 2018: Cell Phone Location Data for Travel Behavior Analysis. Washington, DC: *The National Academies Press*. doi.org/10.17226/25189.

Shiomoto, K., et al, 2023: A novel network traffic prediction method based on a Bayesian network model for establishing the relationship between traffic and population. *Annals of Telecommunications*. doi.org/10.1007/s12243-022-00940-9.

Tarkaa, N.S., Mom, J.M., Ani, C.I., 2011: Drop Call Probability Factors in Cellular Networks. *International Journal of Scientific & Engineering Research Volume 2, Issue 10.*

Tayal, S., et al, 2017: Site Suitability Analysis for Locating Optimal Mobile Towers in Uttarakhand Using GIS. *38th Asian Conference on Remote Sensing – Space Applications*: Touching Human Lives, ACRS 2017.

Wang, Q., et al, 2020: Optimizing the ultra-dense 5G base stations in urban outdoor areas: Coupling GIS and heuristic optimization. *journal of Sustainable Cities and Society*. doi.org/10.1016/j.scs.2020.102445.