The impact of ground-surfaces on the traffic mobile using statistical methods

A. Chariete, O. Baala, A. Caminada
IRTES – SET Institute, UTBM 90000, Belfort, France
{abderrahim.chariete, oumaya.baala, alexandre.caminada}@utbm.fr

***Abstract*—Cellular phones can be used as indicators of the customer's location using the detection of presence of the mobile phones during a day time in urban environment. An urban environment is composed of intrinsically different types of ground. In this paper, we study the impact of ground-surface types on the traffic mobile using statistical methods. Results show that the traffic is greatly influenced by the ground-surfaces covered by the cells.**

***Keywords*—traffic mobile; ground-surface; statistical methods.**

I. Introduction And Related Work

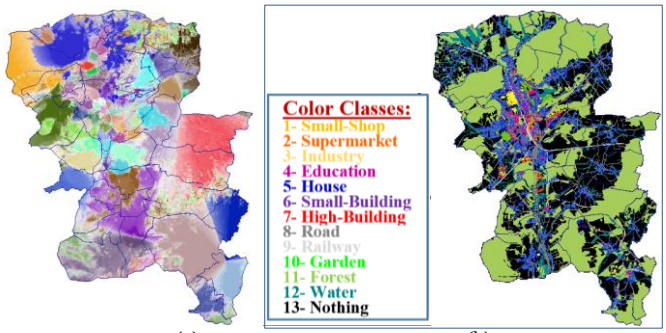
The improvement of Quality of Service (QoS) in mobile networks is a major topic for network operators. In this context, several studies [1,2,3] have shown that to improve the performance of network, data analysis must be applied to construct scenarios from data reported from the network. The exploitation of such scenarios allows developing adapted optimization algorithms and implementing corrective actions to improve the QoS and increase the network performance. The measurements and the Key Performance Indicator (KPI) acquired from a network provide information about network operation and performance. Indeed, the KPI have a range of indicators that cover various aspects of network performance in terms of traffic, call-drops, interference, handover, etc. The measurements, the counter-analysis and the claims of subscribers are information that will allow analyzing and detecting problems in the network.

A statistical learning approach for extracting a model from data with applications to self-organizing network in Long Term Evolution (LTE) functionalities has been proposed in [2]. The proposed model provides closed form expressions that approximate the functional relations between the KPI and the Radio Resource Management (RRM) parameters. Since models allow anticipating the behavior of a network sub-system to new values of RRM parameters, regression analysis methods have first been used to establish the functional relations between the handover parameter and the KPI. It has been shown how the obtained model can be used for monitoring and auto-tuning the network.

In this paper, the results of different statistical analysis methods applied to cellular and ground-surfaces data are presented. The aim is to highlight correlations between the ground-surfaces types and the traffic data. This approach can serve as a management tool of territory, urban network transport, or pedestrian displacement. The paper is organized as follows: In Section 2 we describe the coverage and the ground-surface data used in our study. In Section 3, we present the results of multiple linear regressions that show the correlation between the cells regarding the traffic data. In Section 4, we present the results of factorial and clustering methods and discuss their interpretation. Finally, in Section 5 we present the conclusions.

II. Description Of Data

The analysis needs to combine data from different domains and sources (coverage data and ground-surfaces data). This Section describes how we got the data and gives insight about data quality. In our analysis, we have considered the coverage data originating from one real GSM/UMTS network.



(a) (b)

Figure 1 : a) coverage areas of cells of the GSM/UMTS network, b) classification of ground-surface

A. Coverage Data

The coverage data used in our analysis are collected from 148 cells over a 20kmx40km zone, as shown in Figure 1-a. Each color represents the coverage area of one cell. After cleaning these data, only 67 cells were kept for our study. Among the multiple measured variables, we used calls data, such as on the number of Incoming-Calls (*IC*) and the number of Outgoing-Calls (*OC*). These data were acquired during full days from 6:00 am to 11:00 pm, with a cumulus of *IC*/*OC* every quarter hour. For each day, we had one matrix for each *IC* and *OC* of 67 lines (cells) and 71 columns (quarter hour). Each matrix coefficient (*i*, *j*) represents a number of *IC*/*OC* from/to the cell *i* and cumulated during the quarter hour *j*.

B. Ground-Surface Data

Mobile phones can be used as indicators of the customer's location using the detection of the presence of the terminal on the network coverage. Firstly we have to recognize the ground-surface of the area before associating the mobile location to a human activity. The characterization of the geographical area is represented by the following ground-surface classes: *Small-Shop*, *Supermarket*, *Industry*, *Education*, *House*, *Small-Building*, *High-Building*, *Road*, *Railway*, *Garden*, *Forest*, *Water* and *Nothing* for any other unclassified area. Figure 1-b shows the distribution of those classes on the map of the studied area.

In our analysis, the *Forest*, *Water* and *Nothing* classes are not considered for mobile location as the load an insignificant amount of traffic. A grid of 25mx25m is applied on the studied area and we have computed the number of pixels for each ground-surface category for the coverage of each cell. The resultant matrix dimension is 67x10 corresponding to 67 cells and 10 categories of ground-surfaces. Each matrix coefficient (*i*, *j*) represents a number of pixels of ground-surface category *j* and covered by the cell *i*.

III. Correlation Between Cells Regarding The Calls Data

The linear regression method was chosen for its simplicity and our experience on this problem. It can be applied directly to verify the existence of relations between calls data. To perform this analysis, we sought to highlight the presence or absence of relations between different cells regarding the calls data, *IC-IC* and *OC-OC*, using 2D graphics, and to interpret them. Given the large number of cells, the dimension of the correlation matrix is 67x67 and it is not possible to present all the correlation coefficients in this paper; a sample of this matrix is shown in Figure 2 with some significant cases.

Base Station	11150	11105	11106	11182	11152	11168	11172	11134
11126	0,377	0,722	0,764	0,629	0,695	0,771	0,644	0,560
11148	0,360	0,462	0,595	0,344	0,584	0,567	0,600	0,460
11149	0,479	0,220	0,210	0,075	0,271	0,231	0,304	0,425
11150	1	0,476	0,374	0,269	0,430	0,404	0,504	0,457
11105	0,476	1	0,693	0,503	0,737	0,765	0,674	0,623
11106	0,374	0,693	1	0,750	0,876	0,896	0,768	0,671
11182	0,269	0,503	0,750	1	0,745	0,741	0,669	0,470

Figure 2: A sample of the matrix of correlation regarding the *IC*

In this sample, the highest correlation between cells relative to the *IC* belong to one day was found between the cells 11168 and 11106, with a correlation coefficient equal to 0.896. The scatter plot of this correlation is shown in Figure 3. This scatter plot is giving the number of *IC* of cells 11106 and 11168 in the *X* and *Y* coordinates respectively. The points are linearly scattered along the regression line. This means that these cells are linearly related (strongly correlated regarding the *IC* with a correlation coefficient equals to 0,896). The cell 11106 covers mainly: *Small-Shop* (28%), *Supermarket* (10%), *Small-Building* (15%) and *Road* (30%). Figure 4-a recapitulates these information. Figure 4-b shows that the cell 11168 covers *Small-Shop* (21%), *House* (15%), *Small-Building* (22%) and *Road* (27%). Consequently, on this sample we may do the hypothesis that both cells have a similar behavior regarding the *IC* because they cover very similar ground-surfaces. Because of high attendance of mobile customers in these types of ground-surfaces, the likelihood of calls increases.

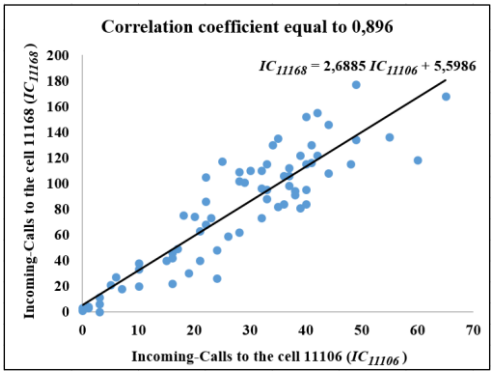
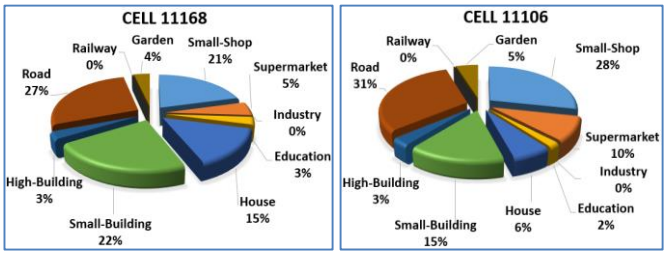


Figure 3: Scatter plots of *IC* between cells 11168 and 11106.



(a) **(b)**

Figure 4: Ground-surfaces distribution of a) cell 11168 and b) cell 11106.

IV. Correlation Between Cells Regarding The Ground-Surface

In our study, the Principal Component Analysis method (PCA) [3,4] is used to compute the correlation of cells regarding the pixel ground-surface on their own coverage. A projection of the results is given for the two main axes of the PCA in Figure 5. The projection of the cells (blue) and ground-surfaces (red) on a 2D plane is formed by the first two principal factors (*F1* and *F2*) with more than 56% of understanding.

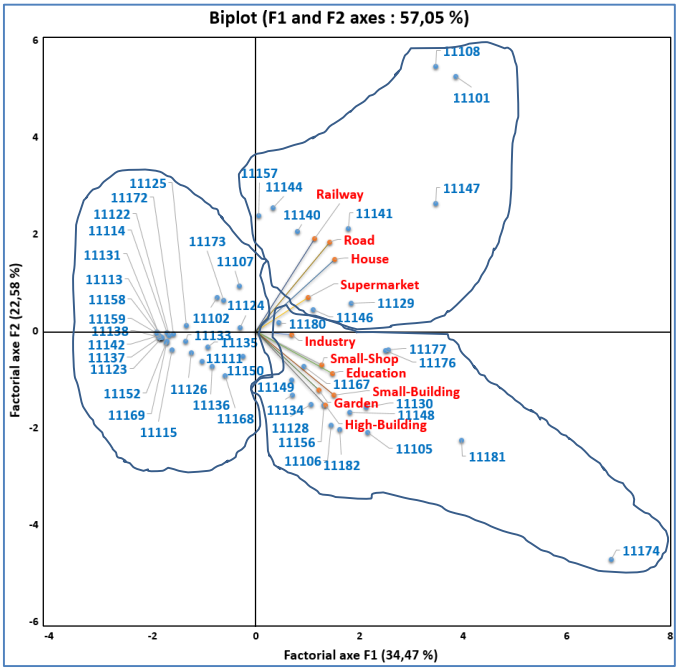


Figure 5: Scatter plot of cells correlated regarding to the ground-surface.

This 2D projection allows us to visually identify the groups of cells covering similar ground-surfaces, and the cells which have singular features. It shows that all cells are divided into three groups. To see more accurately the influence of the ground-surface on the traffic mobile, we identify accurate groups of cells correlated regarding the ground-surface by the use of clustering methods. The Hierarchical Ascending Classification (HAC) [5] and the *k*-means [6] methods were used for this purpose. The projections of cells on principal factors were considered as new variables, and then HAC was launched to see the hierarchical tree (Figure 6). The resulting Dendrogram identifies three classes of cells, shown under the horizontal line. Then, the optimization by the *k*-means algorithm was deployed with *k* equals to 3. By this way, a clear segmentation was obtained, combining cells that cover similar ground-surfaces, and separating those covering different ground-surfaces (Figure 7).

To evaluate this classification obtained by *k*-means, the Silhouettes algorithm was applied [7]; consequently three misclassified cells were detected. These cells were reclassified and the average of obtained Silhouettes index S_i has been improved from 0.55 to 0.57.

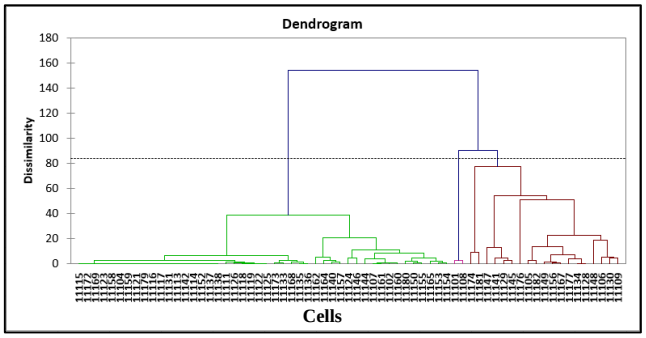


Figure 6: Classification of cells by HAC regarding to the ground-surfaces.

Class	1	2			3
Objects	12	42			13
Sum of weights	12	42			13
Intra-class variance	16,304	1,886			12,394
Minimum distance to centroid	1,691	0,596			1,402
Mean distance from centroid	3,579	1,200			3,075
Maximum distance to centroid	6,905	3,151			6,346
	11147	11173	11137	11164	11148
	11140	11142	11138	11104	11149
	11141	11102	11111	11179	11105
	11101	11122	11131	11118	11106
	11144	11123	11133	11119	11182
	11146	11124	11169	11121	11134
	11129	11125	11180	11161	11156
	11176	11126	11113	11165	11128
	11108	11150	11114	11116	11130
	11157	11152	11115	11117	11174
	11145	11168	11107	11153	11177
	11162	11172	11158	11154	11167
		11135	11159	11155	11181
		11136	11160	11109	

Figure 7: Classes of cells regarding ground-surfaces, obtained by applying the *k*-means clustering method.

V. Impact Of Ground-Surface On The Traffic

After getting a stable classification of three classes of cells regarding the ground-surface, the correlations regarding the traffic were observed in each class separately. As shown in Figure 8, in the first class, 67% of cells were correlated regarding the traffic, whereas only 17% were correlated in the second class, and 12% in the third class. This Figure is one sample of results for one day (that was Monday).

For more relevance in the interpretation of the impact of the ground-surface on the traffic, specific time slots were targeted in specific days. We chose days that haven't similar behavior regarding the traffic, for example Monday, Thursday, Friday and Saturday. For each day, we chose the peak hours of traffic as follows: (7:30 am - 8:30 am), (11:30 am - 12:30 am) and (5:00 pm - 6:00 pm). In addition, we added (10:00 am - 11:00 am) as example of the middle of day, and (4:30 pm - 5:30 pm) where the traffic mobile is maximum, especially in Fridays.

It shows that cells in each class have a similar behavior in various time slots of the day regarding the traffic. These results are similar in each selected day. It is possible that two different types of ground-surface possess similar behaviors of cells regarding to the traffic. It also happens that some cells have a majority type of ground-surface that dominates the traffic.

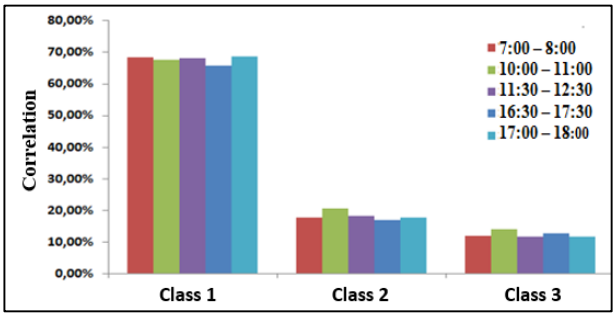


Figure 8: Correlation of cells per class and per time slot regarding to the *IC*.

VI. Conclusion

The results of linear regression, applied to the Incoming-Calls and Outgoing-Calls data collected from one GSM/UMTS network, highlight the relations between cells regarding the traffic data. The PCA applied to the ground-surfaces data gives approximate groups of cells that cover similar ground-surfaces. This result was refines with clustering methods. The global view on these results shows that the influence of ground-surface on the traffic of each cell reaches approximately 67% in the first class, where cells have a very similar ground-surfaces distribution. In the second class, the influence reaches approximately 17% and the cells have little difference types of ground-surfaces. In the third class, it reaches approximately 12% and the cells have singular ground-surfaces distribution. This study confirms that both cells which cover similar types of ground-surfaces and cells which cover predominantly one/many ground-surfaces generate similar traffic. It means that the traffic on the cellular networks is greatly influenced by the types of ground-surfaces covered by the cells.

This approach is destined for the future LTE networks. Operators can benefit from the statistical analysis on today 2G/3G networks to monitor the network resources in function of the ground-surface identity on the current coverage. This approach can also help to understand the urban mobility and the fluidity of the city in order to manage the transport services (e.g. bus stops, lines and timetables).

Acknowledgements: this work was done under a contractual collaboration with Orange Labs.

VII. Bibliography

[1] J. Laiho, M. Kylvaja, and A. Hoglund, "Utilization of advanced analysis methods in UMTSnetworks," Vehicular Technology Conference (55th IEEE-VTC Spring), vol 2, pp. 726 - 730, 2002.

[2] I.T Moazzem, S. Berna, and A. Zwi, "Statistical learning for automated RRM: Application to eUTRAN mobility," Orange Labs, IEEE ICC. France, 2009.

[3] B. Schölkopf, J.Platt, and T.Hofmann, "In-Network PCA and Anomaly Detection," Processing Systems 19: Proceedings of the 2006 Conference, pp. 617 – 624, 2007.

[4] I.S. Lindsay, "A tutorial on Principal Components Analysis," February, 2002.

[5] G. Gruvaeus, and H. Wainer, "Two additions to hierarchical cluster analysis," British Journal of Mathematical and Statistical Psychology, 25, pp. 200-206, 1972.

[6] E. Alpaydin, "Clustering," Chapter in book: Introduction to Machine Learning, MIT Press, pp. 143 – 162, 2010.

[7] P.J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, 1987.