Consiglio Nazionale delle Ricerche
Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti"

V. Sagarese, A. Bragagnini, G. Felici, C. Gentile, G. Stecca

IDENTIFYING PEAK EVENTS FROM MOBILE PHONE PRESENCE DATA

R. 20-04, August 2020

Valeria Sagarese – Alstom Italy, Italy, email: valeria_sagarese@hotmail.com .

Andrea Bragagnini – TIM Service Innovation, Italy, email: andrea.bragagnini@telecomitalia.it.

Giovanni Felici – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR – via dei Taurini 19, 00185 Roma, Italy, email: giovanni.felici@iasi.cnr.it.

Claudio Gentile – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR – via dei Taurini 19, 00185 Roma, Italy, email: gentile@iasi.cnr.it.

Giuseppe Stecca – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR – via dei Taurini 19, 00185 Roma, Italy, email: giuseppe.stecca@iasi.cnr.it.

This work has been partially supported by MIE program - Intelligent Ecosustainable Mobility (Mobilità Intelligente Ecosostenibile) [Grant No. CTN01_00034_594122]. The data used in this work are provided by TIM and belong to a commercial offer of the company; the examined subset are made available for the activities of the research project.

ISSN: 1128–3378

ISSN: 1128-3378 – “Collana dei rapporti dell’Istituto di Analisi dei Sistemi ed Informatica”

Research Report Series of

Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti” – CNR

via dei Taurini 19, 00185 ROMA, Italy

tel. ++39-06-4993-7102/1
fax ++39-06-4993-7106
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

This paper describes a novel procedure for identifying peak events, starting from the data of presence from mobile phones. The aim is to support the identification of aggregation of peoples with respect to time-series of presences in a given territory. Specifically, the data used in the work consist of a database containing the values of presence of mobile cell phone users with a time interval of 15 minutes. The approach used for the peak identification is based on statistical and machine learning (ML) methods. The study describes the procedure used to recognize a combination of percentiles suitable for the recognition of peak events linked to scenarios related to large events of collective interest. Furthermore, to validate this procedure, supervised learning methods are adopted, namely Logistic Regression (LR) and Support Vector Machines (SVM). Test results confirmed that both methods can recognize peak events with remarkable precision when obtained with the optimal percentile combination method.
Key words: Peak Events, Machine Learning, Forecasting, Optimization, Intelligent Mobility

1. Introduction

Intelligent mobility entails the integration of key enabling technologies, such as sensor networks and data mining with innovative organization model and decision support systems. In this context, the analyses described in this work start from the manipulation of a large amount of data, organized in different databases, representing the real values of presence, associated with the mobile phones used for the identification of peak events. The analysis is done in order to identify a technique of recognition capable of predicting the areas in which people is more concentrated, with respect to events occurring in a large area. The use of presence data associated with mobile phones make it possible to identify the peak events of collective interest and supporting the management of the event.

In the literature several studies can be found where the analysis is conducted on the basis of data derived from mobile telephones, whose purpose is to record the user's positions in consecutive calls, as in [3], or the traffic estimate for the management of incident events, as in [9]. Several studies have been conducted on the basis of GPS data with the aim of investigating a series of important social phenomena, including the heterogeneity of the activity spaces, the dynamic nature of spatial segregation and the contextual dependence of subjective wellbeing [7]. In other cases the digital traces of a certain number of samples within a certain temporal and spatial interval are analysed, with the aim of highlighting the distribution of distances and waiting times between consecutive positions,[1, 8]. However, with the respect to the identification of areas where user activities are concentrated, a few studies can be found in literature, only using the monitoring of the movement of people, so considering known traffic flows [6].

In the work of [10], an AdaBoost based learning procedure is applied to a problem of rare event detection. In this case the authors work on a time - space dimension problem where the rare events are in fact faces to be recognized from a monitoring system. The task of finding anomalies under some conditions, can be viewed as similar to the one of finding outliers. In this case, it is worth reading the work of [11] where they propose a measure based on the "holo-entropy" and optimization method to find outliers. Their approach falls in the category of unsupervised methods. Under the same category fall the work of [2], where they propose a clustering algorithm and a rule based methods to detect anomaly treated as outlier.[5] use markov chains in order to detect anomalies in spatio-temporal series, using an approach close to the signal processing. Another interest work, whose scope is close to ours, is the one from [4], where a density based approach is used in order to detect anomalies in time series in the case where the data refer to GPS sensor.

In the present work, the identification of peak events is carried out starting from the raw presence data, and therefore without any information regarding the flow distribution or the position taken by the user. Moreover, we propose methods to 'tag' the data which enable the use of supervised based approaches in a cascade fashion.

The dataset used to prove in a real case the effectiveness of the approach refers to the "Salone Internazionale del Mobile" in Milan, Italy. This is the biggest fair event in the city and one of the biggest event for furniture and fashion industries in Europe. The main exibition is located in the dedicated fair area of Rho city. During the fair days several events are organized in the city center. It is indeed of interest to have an automated procedure in order to individuate peak events associated to the organization of fairs, and inside the fair days to localize the most attracting places. This information is useful in particular for organizers and planners of the events, in order to understand how to size the resources allocated to the event during the fair. The information can be useful also to the local administrator for pre allocate resource for crowd management. The method is applied to a dataset where presence data are generated based on user's counters allocated for cell, collected over 15 minute intervals. The elaboration is made using, as input data, the covering maps of cells for the TIM company's mobile network. The data provided by TIM belong to a commercial offer of the company and the examined subset are made available for the activities of the research project MIE (Mobilità Intelligente ecosostenibile - Intelligent ecosustainable mobility) of which this work is part of. The results are organized in a file which contains the estimation of the number of users connected to mobile's network on the overall province of Milan, which is subdivided in areas of dimensions 140x130 metres. The data is structured in form of matrix characterized by 389 rows and 511 columns, whose elements represent an area of the selected territory, and whose values represent the number of users allocated to TIM's net. The paper is organized as follows. In Section 2 percentile base detection of peak events is presented; Section 3 describe the procedure used to detect the threshold values; Section 4 describes the detection of peak areas through supervised learning methods, and Section 5 conclude the paper.

2. Percentile based detection of peak events

Given a dataset of values representing a specific variable observation, the percentile represents the value below which it occurs a certain percentage of observations. For examples the 10° percentile is the value below which the 10% of observations can be found: in other words, given n values, the 10° percentile is a number such that $(n/10)$ of the values are lower or equal to it. In general, to calculate the k^{th} percentile, if one considers a sample of n values, increasingly ordered, the k^{th} percentile index I_k is given by:

$$I_k = \left\lceil \frac{n_k}{100} \right\rceil \tag{1}$$

The method is called 'nearest-rank' and it can be used with a linear interpolation between the two data with indices equal to the one before and after Ik. The concept is straightforward but when dataset are multidimensional things could be more complicated. In particular, when dealing with spatial data representing events occurring in a specific time horizon we have two main dimensions, and over these two dimension we may have several features describing the data. With respect to the percentiles, which can be used to detect anomalies, the proposed procedure foresees to define two percentile based thresholds. The first one is a spatial threshold, named σ_t, acting as a spatial filter and defined over the time interval t, while the second one is a temporal threshold $\tau_{i,j}$, acting as a temporal filter for each spatial cell (i, j) where i is a measure of the latitude and j is a measure of the longitude. More formally we define x_{tij} the value of presence corresponding to cell of i^{th} row and j^{th} column, at the day t of month related to year considered, at hour fixed; then, for a single cell to be identified as a peak the following conditions must be satisfied:

$$x_{tij} > \tau_{ij} \tag{2}$$

$$x_{tij} > \sigma_t \tag{3}$$

In the addressed test case, the detection of peaks of presence, associated to anomaly events, is made at fixed hours for all days of February, March, and April concerning data belonging

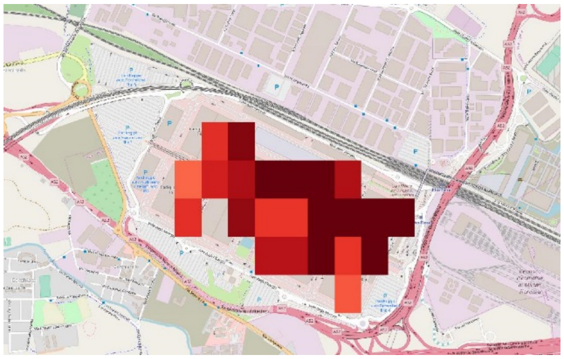

Figure 1: Rho peak areas on April 5, 2017 at 10:00

respectively to 2016 and 2017. In particular, at fixed year, month, and hour, the data of presences concerning the interested zone is extracted, for each day. Subsequently a combination of threshold values is defined in order to filter presence data and allowing to consider as peak values only those who are higher to it, thus discarding the remaining presence values. Thus, the definition of threshold value is made by using the concept of percentile.

The first step of the procedure foresees the tuning of the threshold which bound values are case specific. In this phase the setting is done also by a visual help provided by the hit maps of the aggregations of people during past events in locations and by numerical check on the number of peaks filtered. Depending on the application case this phase help filter false positives and limit the number of desired peaks. In some sense the number of desired peaks is a priori decision which can be set by the decision maker. After the peak assessment, as described in the follows, in the next section we provide methods which can be used to validate and improve the selection of the threshold through optimization and ML.

The empirical phase of peak identification has been conducted by using the visualization on map of the presence data thanks to GIS (Geographical Information System) technology, and by tuning the threshold values. The thresholds are computed in the area of interest which is partitioned in cells selected by column j and row i. The entire area is represented on a matrix for each time slot t, with the granularity set to 1 hour. The 80° percentile is selected as spatial threshold σ_t, while the temporal threshold $\tau_{i,j}$, has been set to 95° percentile.

Two threshold values are needed because each single cell is provided of its 95° percentile, evaluated on all days of month fixed. So, the limit of 80° percentile allows to remove less significant peaks on each single cell. Moreover, this limit is needed in order to support the purpose of the application case which aim is to highlight only big events without generating too many false positives. Finally, the matrix of peak areas is a matrix having size equal to the area of interest, where each single element is equal to the value of presence if the limit is satisfied and otherwise it is zero.

According to the procedure described above, Figures 1, and 2 reproduce the main fair area of Rho, during specific times in 2017 trade fair when peak are recognized, and where the darkness of the squared overlays color is proportional to the presence value.

However, these values of threshold have been identified empirically, so for this reason, an analytical procedure for searching these values has been proposed and applied in the following of this paper. After obtaining the percentile values, the peak areas have been identified considering two different methods, namely LR and SVM. The identified peak areas have been compared with those obtained with the percentile method.

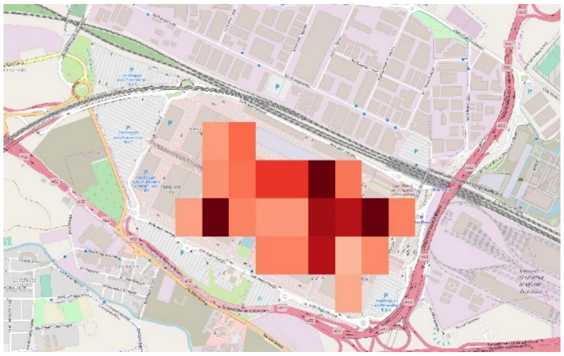

Figure 2: Rho peak areas on April 6, 2017 at 20:00

3. Detection of thresold values

In order to computationally find the best combination of percentiles maximizing the peak events recognition rate, a database has been created for each combination, considering that the percentile functions of space and of time vary in a range between 70 and 100 respectively. Each database has obtained for the year 2016 and 2017 considering the Rho area and on this dataset we assumed the value 1 for the peak and 0 for the non-peak. After setting these value based on the knowledge of past peak events (Rho fair in 2016 and 2017) we are able to find the best threshold combination.

For each assigned combination, the number of peaks on the total data contained in each database was evaluated and, among these, those combinations that define the lowest number of peaks were considered in order to identify only those areas in which actually the peak of presence is recorded, avoiding to consider less significant areas. Following this principle, we could leave the time threshold set to 95° percentile both for the year 2016 and for the year 2017, while we focused on finding the best space threshold in the range between the 70° and the 95° percentile.

For the space percentile, a further consideration can be made. Plotting the presence data against the time, we are able to know the timeslot with the highest presence during the peak event time horizon. In particular, for the two analysed datasets, the greatest number of visitors occurred between 13 and 14 April 2016 and between 5 and 6 April 2017. As a consequence, the total number of peak events identified by each combination can be compared with those identified only in the days when there are more entrances to the show. In this way we can test the effectiveness of the threshold levels (and the proposed ML approaches) to replicate the manually identified peaks. The procedure can be clarified by Figures 3, and 4, showing the trend of presences on Rho for the year 2016 and 2017 at 12:00 is shown:

The optimal combination is the one identifying the maximum number of peaks and at the same time maximizes the percentage of peaks in days of the fair of the year 2016 and of the year 2017 in which it is attested a high presence rate. The threshold performance can be measured with the respect to the number of recorded peaks during the event, and with the respect to the percentage of peaks recorded during the event on the total of the recorded peaks, leading to eventually non dominated combinations.

For the year 2016 two best, non dominated combinations are obtained, the one for which the

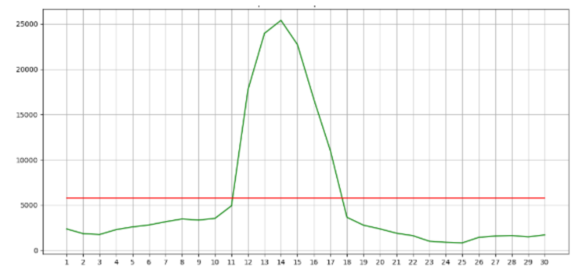

Figure 3: The trend of total presence of Rho in the month of April 2016 at 12:00 o'clock

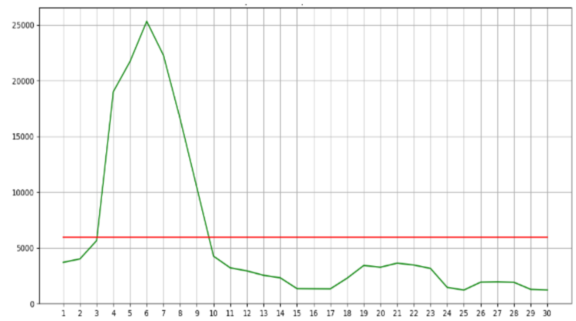

Figure 4: The trend of total presence of Rho in the month of April 2017 at 12:00 o'clock

time threshold is equal to the 95° percentile and the spatial one is equal to the 70th percentile; the second one with time threshold is equal to the 95° percentile and the spatial threshold equal to the 95° percentile. For the year 2017, instead, there are three non dominated combinations, which namely are: *i*) the one for which the time threshold is equal to the 95° percentile and the spatial one it is equal to the 70° percentile; *ii*) the one for which the time threshold is equal to the 95° percentile while the spatial threshold is equal to the 80° percentile; and *iii*) that for which the time threshold is equal to the 95° percentile while the spatial threshold is equal to the 95th percentile. The combination 0.95, 0.80 is not considered dominated by the 0.95, 0.95 because their accuracy is very close and over 86%. The detailed results are depicted in Table 3 for the year 2016, and in Table 3 for the year 2017.

Table 1: combinations of non-dominated percentiles for the year 2016 (in bold), Fair occurred in the days 13-14 April

time thresold	**spatial thresold**	**peaks during fair**	**total peaks**	**%**
0.95	**0.70**	**690**	**861**	**80.14**
0.95	0.75	597	748	79.81
0.95	0.80	472	592	79.73
0.95	0.85	354	450	78.67
0.95	0.90	243	308	78.90
0.95	**0.95**	**122**	**149**	**81.88**
0.95	1.00	20	24	83.33

Table 2: Combination of non-dominated percentiles of the year 2017 (in bold). Fair occurred in the days 5-6 April

time thresold	**spatial thresold**	**peaks during fair**	**total peaks**	**%**
0.95	**0.70**	**714**	**839**	**85.10**
0.95	0.75	613	722	84.90
0.95	**0.80**	**495**	**575**	**86.09**
0.95	0.85	374	436	85.78
0.95	0.90	250	291	85.91
0.95	**0.95**	**126**	**145**	**86.90**
0.95	1.00	22	26	84.62

4. Detection of peak areas through supervised learning methods

Supervised learning methods allow pattern classification thanks to training set where user knowledge is coded (a set of attributes tagged with the membership class). A supervised learning algorithm then looks for the function assigning to each data its class of membership. Each statistical learning process is divided into two phases:

1. learning phase, where the algorithm analyses training data and recognizes it the similarities in the data to build a model that approximates;

2. test phase, where the model generated during training is tested on a different set of data for real performance.

Therefore, classifying activity consists in identifying which category belongs to a new observed data, based on a data set called training of which the category of membership is known. The combinations of percentiles found against manually identified peak areas can be used to test the effectiveness of supervised ML techniques such as LR and SVM. In this way we are able to classify automatically events. For the specific case, by using the combination of percentiles obtained in the previous section, six datasets were processed in order to test the two ML techniques over the different percentile thresholds.

For each dataset the following 10 features (independent variables) are considered as dataset columns: latitude; longitude; time; value of presence; day; day of the week (1-7), values of the presence in the adjacent north, south, east and west cells with respect to each considered cell. The input dataset is then completed by a last column which stores the values of the dichotomous variable derived from the percentile combination obtained. The six datasets are obtained considering the two analysed years as separated or joint, and considering or excluding the week of the furniture fair (which tend to cover all other eventual events). The resulting datasets are built in the following way: starting from the two datasets of presences in the month of April 2016 and 2017 we define 4 datasets considering or excluding the week of the fair. This is needed because in the datasets without the fair we are able to recognize also peaks not related to the fair, while in the datasets considering the fair, all fair related events are revealed. Two more datasets are built considering together the data of both years with and without the fair week.

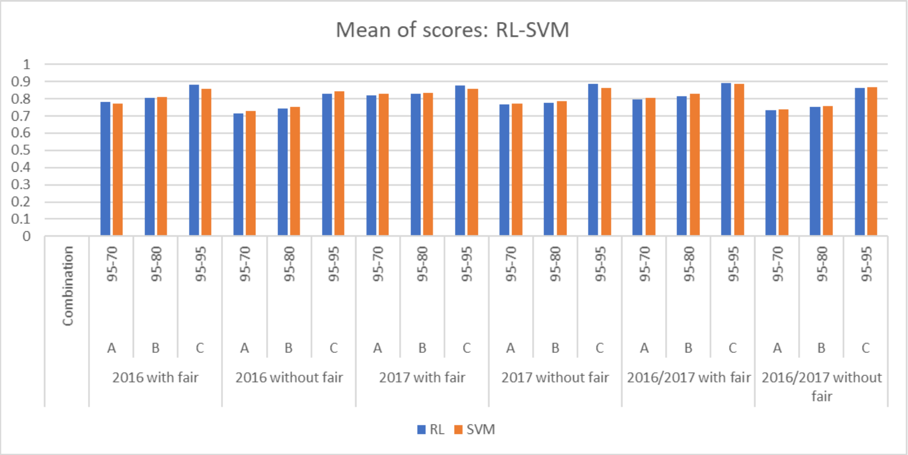

Figure 5: The mean of scores for the Logistic Regression and for the Support Vector Machine

For each dataset a LR analysis and a SVM analysis was carried out. For each of these analyses a validation was made through the cross-validation procedure at 10 folds.

Starting from the six datasets obtained from the percentile combination that maximizes the search for peak areas, a classification analysis was conducted using the LR and the SVM with polynomial kernel of degree 1 and 2, but the results discussed in the following are obtained with the grade 1 polynomial kernel. In a further test set we compared this SVM against the SVM with the grade 2 polynomial kernel. On the two tested kernels, over the given datasets, the second one did not provide noticeable improvement and required more computational time. This is and indication that non linear methods could be useful only for more complex datasets. For both methods (LR and SVM with 1-degree polynomial kernel) a 10 folds cross validation process was carried out. This is a statistical technique involving the splitting of the database into k subsets of same size (also known as k-fold validation) and, at each step, the k-th subset is considered as validation dataset, while the other subset composes the training datasets. Thus, for each of the k parts (usually $k = 10$) the model is trained, avoiding problems of overfitting, but also of asymmetric sampling of the training dataset with respect to the validation one, a typical situation occurring in the splitting of the dataset into only two parts.

For each subdivision of the database it is possible to assess the performance of the used model by considering the mean of the method performance over the 10 trials. Figure 5 reports the average of the scores related to the 10 iterations of the cross-validation, both for the LR and for the SVM, considering the three different non-dominated combinations under study, corresponding to the three bold rows in Table 3 and denoted here with A, B, and C. In detail, A is the case in which time threshold and spatial threshold is 0.95 and 0.70 respectively; B is the case in which time threshold and spatial threshold is 0.95 and 0.80 respectively; C is the case in which time threshold and spatial threshold is 0.95 and 0.95 respectively.

The combination that best describes the model in terms of recognition of peak events is the combination C. This, in fact, allows to filter more so it is more suitable for the recognition of events. With respect to the two ML methods, in particular for the combination C, the two methods are practically equivalent, with a slighter preference towards the LR. A deeper analysis

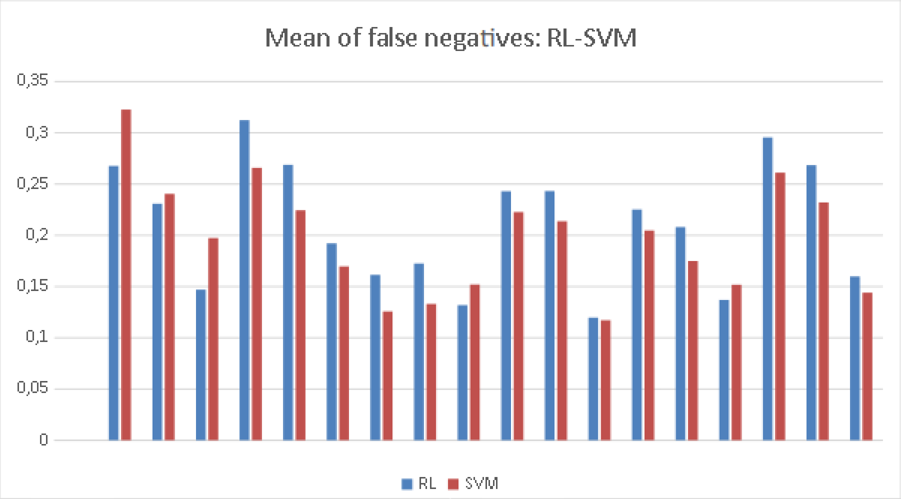

Figure 6: The mean of false positives (as fraction) for the Logistic Regression and for the Support Vector Machine for each of the analysed dataset

can be made looking at the trend of false positives and false negatives for the two methods. The false positive is when the dichotomic variable Y takes the value 0 (not peak) in the training set and value 1 (peak) in the prediction set. On the other hand, the false negative occurs when the dichotomic variable Y takes the value 0 (not peak) in the prediction set and value 1 (peak) in the training set. This implies that the performance of the model is also related to the prediction accuracy of the peaks compared to the training values. For the $0.95 - 0.95$ combination, the method that minimizes the percentage of false positives and false negatives is the LR for the datasets including the week of the fair, while the SVM appears to have lower false positives and false negatives for the datasets where the fair week has been excluded. The results are showed in Figures 6 and 7.

Comparing the mean cross-validation scores for grade 1 and 2, SVMs shows that these values differ by a few units over 1000, how depicted in Figure 7. For this reason, because the second-degree analysis does not provide a relevant improvement of the model in the estimation of peak events, but it requires a significantly higher computational time (200 minutes compared to 55 seconds for the analysis of degree 1 (on a 8GB i7 5600U Windows 10 machine, with the scikit-learn python library) is enough to refer to the analysis with a polynomial degree 1 kernel. The values relating to the analyses just mentioned are shown in Figure 8.

5. Conclusion

In this paper we have proposed a method for the search for a percentile combination capable of recognizing peak events during an event of considerable collective interest, with no information related to traffic flows or to the positions taken by the user instant by instant. The proposed approach integrates the classical empiric set of threshold bound based on percentiles; in fact, the approach propose an assessment of the thresholds which compare different combinations of percentiles and automatize the selection by use of ML. This approach is useful in consideration of the fact that the data have multiple dimensions. In particular, in the considered application case, the dataset has a spatio-temporal dimension. Moreover, in the machine learning phase, data could be enriched with features detailing the time and the space. In the tests we compared two

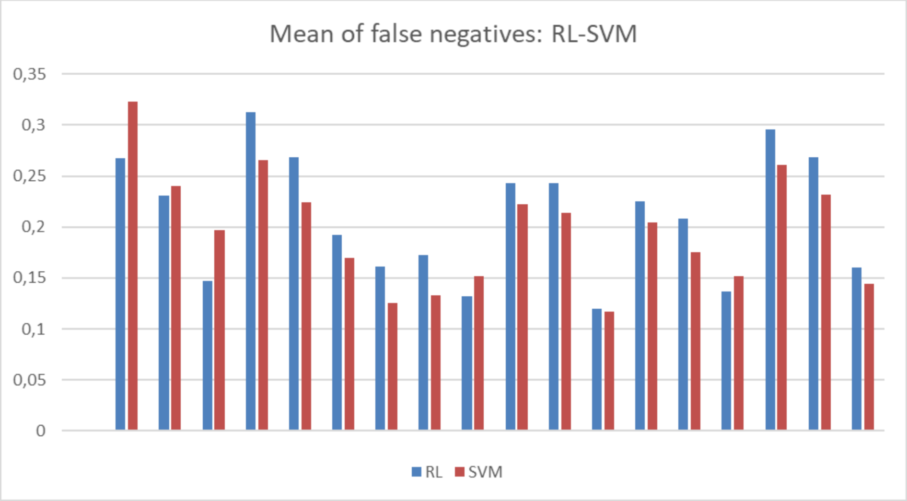

Figure 7: The mean of false negatives (as fraction) for the Logistic Regression and for the Support Vector Machines for each of the analysed dataset

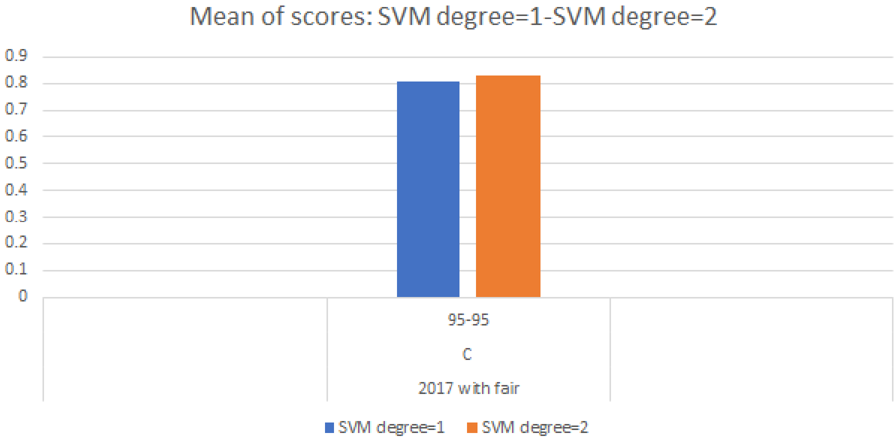

Figure 8: The mean of scores for the Support Vector Machines of degree equal to 1 and 2.

different supervising learning methods resulting in almost equivalent very good performances. We proved that linear models (such as LR and SVM with 1-degree kernel) are sufficient to guarantee very good results in peak detection. The approach validated in this article can be used as a basis in different situations where, with more rich data, a more sophisticated learning model could be assessed.

References

[1] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli, "Multi-scale spatio-temporal analysis of human mobility," *PloS one*, vol. 12, no. 2, p. e0171686, 2017.

[2] P. K. Chan, M. V. Mahoney, and M. H. Arshad, "A machine learning approach to anomaly detection," tech. rep., 2003.

[3] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[4] Z. Liu, D. Pi, and J. Jiang, "Density-baed trajectory outlier detection algorithm," *Journal of Systems Engineering and Electronics*, vol. 24, no. 2, pp. 335–340, 2013.

[5] Y. Meng, M. H. Dunham, F. M. Marchetti, and J. Huang, "Rare event detection in a spatiotemporal environment," in *2006 IEEE International Conference on Granular Computing*, pp. 629–634, IEEE, 2006.

[6] D. Orellana, A. K. Bregt, A. Ligtenberg, and M. Wachowicz, "Exploring visitor movement patterns in natural recreational areas," *Tourism Management*, vol. 33, no. 3, pp. 672–682, 2012.

[7] J. R. Palmer, T. J. Espenshade, F. Bartumeus, C. Y. Chung, N. E. Ozgencil, and K. Li, "New approaches to human mobility: Using mobile phones for demographic research," *Demography*, vol. 50, no. 3, pp. 1105–1128, 2013.

[8] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using wifi signals," *PloS one*, vol. 10, no. 7, p. e0130824, 2015.

[9] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities," *GeoJournal*, vol. 78, no. 2, pp. 223–243, 2013.

[10] J. Wu, J. M. Rehg, and M. D. Mullin, "Learning a rare event detection cascade by direct feature selection," in *Advances in Neural Information Processing Systems*, pp. 1523–1530, 2004.

[11] S. Wu and S. Wang, "Information-theoretic outlier detection for large-scale categorical data," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 3, pp. 589–602, 2011.