Understanding students’ mobility habits towards the implementation of an adaptive ubiquitous platform

João Casal
Centro Algoritmi
Universidade do Minho
Portugal
joaocasal@dsi.uminho.pt

Guillermina Cledou
HASLab INESC TEC
Universidade do Minho
Portugal
guillecledou@gmail.com

ABSTRACT

Adapting technological environments to users is a concern since Mark Weiser launched the concept of ubiquitous computing and, in order to do that, is necessary to understand users’ characteristics. In this context, the purpose of this paper is to present a study about students’ mobility habits within a university campus, having the intention of getting insights towards the best place to set an interactive public display and of predicting the main characteristics of the audience that will be present on that spot in forthcoming periods. Thus, the envisioned results of this work will allow the adaptation of the contents exhibited on the device to the audience. To perform the study, a set of logs of accesses to the university’s Wi-Fi was used, data mining techniques were implemented and forecasting models were built, using the line of work suggested by the CRISP-DM methodology. As result, students profile were built based on past wireless accesses and on their scholar schedules, and three time series models were used (Holt-Winters, Seasonal Naive and Simple Exponential Smoothing) to predict the presence of students on the envisioned spot in future periods.

Categories and Subject Descriptors

H.4.2 [Information System Applications]: Types of Systems - *Decision Support*; H.2.8 [Database Management]: Database Applications – *Data mining*; I.5.2 [Pattern Recognition] Design Methodology - *Classifier design and evaluation*

General Terms

Algorithms, Measurement, Experimentation, Human Factors

Keywords

Adaptive Business Intelligence; Ubiquitous Environments; Interactive Public Displays; Prediction Techniques; CRISP-DM.

1. INTRODUCTION

Humans’ mobility has been studied for decades because of its importance regarding aspects like urban planning [13], traffic forecasting [5] and estimating the spread of viruses [11]. Nowadays, emerges the proliferation of personal mobile devices and the availability of Wi-Fi access points on several urban environments. This fact, in one hand, assists the study of the mobility of the persons connected to a network [2], and in the other, enables the transformation of public spaces in adapted ubiquitous environments or, in other words, environments where the technology is adapted to the user in a way that it is transparent and natural [12]. Therefore, understanding which persons are in a given spot at a given period allows the adaptation of the environment, mainly digital situated artifacts like public displays, to the bystanders’, fostering benefits in fields like advertisement, education or leisure.

In this work it is considered the mobility of students on a university campus, based on Wi-Fi access logs. The objectives are to get insights towards the grounded implementation of an interactive public display in the campus in terms of what is the best location for it, and to understand if it is possible for the artifact to predict the characteristics of the students present and adapt the contents exhibited to the expected inconstant audience.

It is understandable that when implementing an interactive artifact like the one intended the main goal is the engagement of the audience towards interaction [7]. In the specific case of study, the interaction with the envisioned public display will be made through personal devices like smartphones or laptops and, consequently, the Internet connection is essential. Therefore, the university network of access points (AP) called EDUROAM is the key aspect that enables students interaction with the ubiquitous platform and, through its logs, assists the study of their mobility inside the campus in order to foster the adaptability of the system.

2. MATERIALS AND METHODS

2.1 Adaptive Business Intelligence and CRISP-DM

Accordingly to Michalewicz, Schmidt, Michalewicz, & Chiriac [8] the term Adaptive Business Intelligence can be defined as “the discipline of using prediction and optimization techniques to build self-learning ‘decisioning’ systems” and it includes elements of data mining, predictive modeling, forecasting, optimization, and adaptability. Thus, it is believed that this big umbrella of techniques adapts exactly to the business issue specified before. In this case, it will be necessary to apply data mining on the logs of EDUORAM to find patterns of presence in specific spots of the campus (identified by its APs) and to profile students through its network accesses. From this first task will outcome the best spot for placing the interactive public display and will also result prepared data to be used in forecasting technics, the second task, which will have the goal of estimating which students will be at the selected spot in forthcoming periods. As future work, the forecasted data (expected students present near the display at each moment) may feed the system optimization algorithms, which will maximize the adaptation of the contents available to the audience that is likely present, turning the platform into a “’decisioning’ system” (the application shall decide which contents to exhibit). However, in this paper the process will end with the forecast of the students present on the selected spot of the campus.

To achieve the underlying goals, the presented study followed the CRISP-DM data mining methodology [3], which suggests the cycle of phases described by Figure 1. As shown, the process evolves in a research-action process around the collected data, where the understanding of the business and of the data is interrelated and may change with the results of the experiments, giving rise to new cycle iteration.

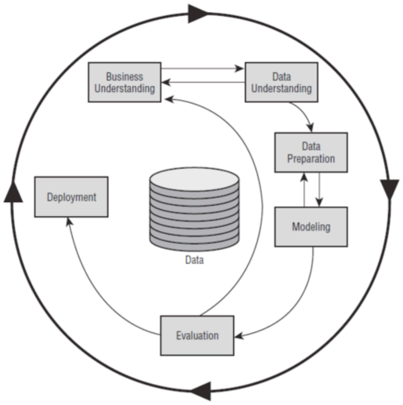

Figure 1. Overview of the lifecycle of a data mining project oriented by CRISP-DM methodology [3]

In this paper will be presented one iteration of the aforementioned cycle, from the business understanding to the evaluation. However, even though it has been carried out only one iterance of the main CRISP-DM methodology cycle, that process is not linear or straightforward. As generically stated before and as Figure 1 shows, there are several inner cycles in the path: data understanding also affects the business understanding, the data preparation also influence the data understanding (this aspect is not in Figure 1 but it occurs) and the modeling normally compels to extra work of data preparation. Moreover, lifting slightly the veil of the conclusion, as the results of the evaluation served the goals of the project, there was no need to start a new iteration of the cycle. An inference that may be assumed and will be verified afterward is that although the well-defined phases of the methodology, its inner cycles make the connections almost seamless. Consequently, it is considered a good approach to present the phases in an integrated manner.

2.2 Business understanding

Given that part of the business understanding was presented on the introduction, at this point the information will be consolidated and systematized.

Business goals. As stated before, the general goal is the introduction of an interactive public display on a university campus. As specific objectives, emerge the following three: (1) determine the best spot in the campus for the display; (2) estimate who is present on that place at each moment; and (3) as future work, maximize the interaction with the display through the adaptation of the contents to the passerby’s or bystanders. For evaluating the accomplishment of the business goals that are subject of this study ((1) and (2)), the success criteria will be (a) the grounded achievement of a spot on the campus to situate the display; (b) a basic identification of the profile of students, specifically their university courses; and (c) an accurate estimation of the number of accesses of students of each course on the selected spot at forthcoming periods.

Situation assessment. The resources needed to perform the study are the logs of accesses to EDUROAM APs, classroom schedules and maps relating the schedules to the positioning of the APs.

Regarding possible constrains that might emerge, typically ethical issues about privacy are highlighted when studies try to understand the mobility of persons [9]. However, in this case, this was avoided because individuals were never identified. The only profiling made is the relationship between MAC addresses of personal devices and university courses. Other constrain that arose, was that for this study were only available schedules of the classrooms and maps with APs of the Department of Informatics (DI) of the university. This constrain reduced the courses that could be identified however, it did not affected the objectives established because if the study could be applied to courses of one department, analogously could be performed to the other ones.

The highest risk for the project success was not being able of mapping MAC addresses to university courses through the access of students to APs when they were in classrooms. This was a key part of the project, because if it were not possible to find the profile of the students the system would not have anything to forecast or to adapt to. The main threats regarding this task were the relationship of N to 1 between classrooms and APs, and the passerby’s through classroom corridors with devices Wi-Fi set on. The actions taken to avoid this menaces will be explained during the presentation of the data preparation phase however, in order for being able to perform the aforementioned part of the project, an assumption had to be made. Because it is possible that classrooms are in the range of more than one AP, it was assumed that the connection of the personal devices was made to the nearest one.

Data mining goals. The data mining objectives may be seen as the project goals transposed in technical terms [3]. Therefore, the aim of selecting the better spot in the campus to place the public display in order to maximize the interactions may be defined as the public spot with a history of more accesses to EDUROAM. This is a good assumption because the mechanisms of interaction with the displays are personal devices, so the place where these have been more used in the recent past is the best for implementing the ubiquitous environment. Regarding the second specific business goal, it can be subdivided in mapping MAC addresses of personal devices to courses and in predicting, based on historical data, which of the identified MAC addresses will be present on the chosen spot at upcoming periods.

As main success criteria for these goals it is reasonable to assume that if experiments confirm that the models implemented predict the moments where students with the profile identified were present in the selected spot, then the aims of the project have been achieved.

Project plan. Defined the goals it is time to set the path to achieve them. Following will be described the available data: the EDUROAM access logs. The analysis of that data, will lead to several insights, from which excel the election of the spot to situate the public display. On the next phase, data preparation, the main steps are building of the student schedules dataset, mapping of courses with MAC addresses and constructing the time series with the accesses of the MAC addresses identified on the spot selected. On the next phase, modeling, the forecast techniques will be chosen accordingly to the characteristics of the time series obtained. These will be applied and their results will be assessed. At the final phase, evaluation, the attainment of the project will be assessed.

2.3 Data understanding

As previously mentioned, the main data source used is the set of logs of accesses to the APs of EDUROAM. Each log have the following information about the connection: date and time, type of access (start or stop), session ID, time of access, mac address of the client, identifier of the AP and number of octets sent and received. At first sight it was clear that the main characteristics of interest of the logs, for the goals established, were the date and time, the mac address of the client and the identifier of the AP. For the study it was granted access to one month of logs to all EDUROAM (accesses on the entire campus), which was equivalent to approximately 8 million of registries. This data was sufficient to understand which would be the best spot to situate the interactive public display. Using a computational tool (R) to analyze this information, it was possible to see that the two APs of a bar in a central building of the university had almost 100.000 accesses on the month available, being this the public spot with more accesses. So, this bar was selected as the best option to situate the interactive public display.

In terms of data quality, as the work is made with automatically created logs, there were no errors identified. However, for the endeavor established, was identified the need previously mentioned of creating a dataset with schedules of courses in order to map the personal devices of students with their courses (through their access to EDUROAM during the classes). The schedules and the maps of APs gathered were, as explained before, from the DI: maps of the two floors of classrooms and schedules of 15 classrooms. With this information was built the schedules dataset composed by the following information: classroom, day of the week, starting hour, ending hour, course and nearest AP. The merge of this dataset with the logs will be explained on the data preparation stage.

3. EXPERIMENTS AND RESULTS

3.1 Data preparation

This phase has the purpose of preparing the collected dataset to be used as input for the modeling phase. To that end, this subchapter is divided in two iterations. The first describes the selection of relevant instances and attributes, the construction of temporal datasets, including one with schedules for a particular set of courses, and the identification of those courses through a particular set of accesses. The second iteration describe the selection of relevant MAC addresses from the dataset obtained in the first iteration, the selection of the instances of accesses that correspond to the APs selected for forecasting and the construction of the datasets that will be used as inputs for the time series algorithms.

3.1.1 First iteration

Instance Selection. The dataset used on this task is the whole dataset of logs of accesses, which consists of almost 8 million instances. Since approximately half of them correspond to "Start" connections and the other half to "Stop" connections, one of these subsets can be removed because it is just needed to know that a connection occurred at that particular spot.

In order to map MAC addresses with courses it is necessary to build a dataset relating the courses schedules with the nearest APs of the classrooms were the courses are lectured. Given that the maps of APs and the schedules collected for this work correspond only to the DI, only accesses of that department, and in particular accesses of the APs that are near the classrooms, are considered in order to match them with a course. Thus, it is necessary to select a subset of instances from the subset obtained in the previous instance selection step. This subset contains the accesses from the APs of the DI that are near the classrooms of the courses for which the schedule was provided.

Attribute Selection. Given the goals of the work, it is clear that many attributes presented in the original dataset are not relevant. Therefore, the following attributes were eliminated: access type, session id, session time, octets sent and octets received. The new dataset consists of four attributes: date, time, MAC address and id of the access point.

Data Construction

To build the dataset of the schedules and nearest APs (of DI) 9 different courses were used: ARQUEOL, CCOM, ENGINF, GEOLOG-P, LA, MBINF, MEI, MIEBIOME and MRSC. This dataset was built with the attributes day of the week, course starting hour, course ending hour, course name, and id of the nearest access point of the classroom where the course is lectured.

In order to map an access of a MAC address to a specific course, several conditions must be met: the access and the course must have the same (a) AP, (b) day of the week, and (c) the time of the access must be between the starting and ending hour of a class of that course. From (b) arises the need of adding a new attribute to the (prepared) logs dataset: the day of the week.

At this point an ambiguity issue came forward: one MAC address could be associated with more than one course. This occurred because it was possible that more than one course was happening at the same time and shared the same AP (the same AP was the nearest of more than one classroom that had classes at the same time). Could also happen that the owner of a MAC address, at some moments, passed by (and connects to) APs that were on the time schedule of other course. To handle this, it was assumed that the course attached to a personal MAC address should be the one with higher occurrences of accesses to the APs related to the schedule of that course.

The outcome of this step was a dataset with the accesses from the DI where each record had two new attributes: day of the week and the course of the MAC address.

3.1.2 Second iteration

Data Selection

On the dataset obtained in the previous data construction process the number of unique MAC addresses per course was not representative of the number of actual students of those courses (it was higher). This happened because some of these MAC addresses were probably just passing by an AP when connected, and their owners did not belong to any of the courses. In order to make the numbers more realistic, all MAC addresses that had less than twelve accesses in the whole month, were removed. From the initial list of courses, 7 were obtained from the matching and selection process: ENGINF, GEOLOG-P, LA, MBINF, MEI, MIEBIOME and MRSC. Table 1 shows the number of MAC addresses per course.

Table 1: Number of identified MAC addresses per course

	Unique MAC addresses found
ENGINF	85
GEOLOG-P	3
LA	28
MBINF	90
MEI	163
MIEBIOME	48
MRSC	9

After, it was necessary to identify all accesses that these MAC addresses made during the month on the APs selected for the forecasting task (the APs of the bar envisioned to receive the public display). To this end, all accesses that do not correspond to these APs were removed (dataset 1), so as all records of accesses of MAC addresses that did not correspond to the ones univocally identified (dataset 2). These two premises conducted to two datasets.

Data Construction

Having the two datasets previously specified, the next step was to use the first to identify the course of each access in the second.

With the resulting dataset was possible to build the temporal vectors of observations that would be used as input for the time series algorithms. As these algorithms require accesses to be recorded in uniform intervals, it was necessary to build, for each course, a table of accesses per unit of time, in this case, hours, for the APs of the bar. This means 24 measures per day, for the 31 days of the month of logs used, per course. The outcomes were 7 vectors (one per course) of 744 observations each (24 observations of number of accesses per day for the 31 days).

3.2 Modeling

At modeling phase, three time series techniques were selected, applied and their parameters calibrated to optimal values. After preparing the time series on the previous section, the first task of this phase was to analyze it in order to select the best prediction technique.

3.2.1 Selecting modeling techniques

As previously stated, time series are stretches of values on the same scale indexed by a time parameter [1]. In this study, it was built a set of time series, where each describes the number of accesses to EDUROAM, per course, per hour at a bar of the university. To select the proper forecast techniques was necessary to be aware of the characteristics of the time series created in terms of the trend, seasonality and cyclic properties. Analyzing the sample of Figure 2 it is easy to see that it has a seasonal pattern with the periodicity of a week. For the sake of legibility of the graph, it is worth mentioning that the month analyzed started on a Tuesday and all the hours of all days were represented (day and night).

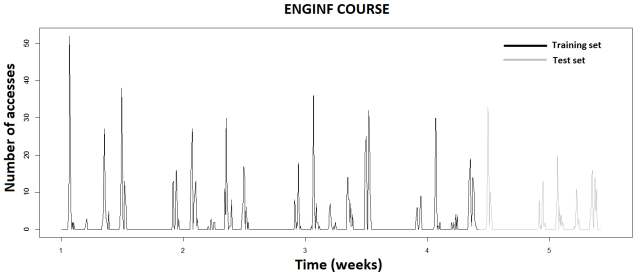

Figure 2: Time series sample of the accesses at the specified spot of an identified course (ENGINF)

After obtaining the characteristics of the time series it is possible to select the appropriate techniques in order to evaluate which provides better forecasts. Therefore, were selected the Seasonal Naïve [6], Holt-Winters [4] and the Simple Exponential Smoothing [10].

3.2.2 Test design, models and assessment

In order to train the time series models and test their performances, it is necessary to separate the observations into a training set and a test set. Five courses were selected to train and test the models: ENGINF, LA, MBINF, MEI, and MIEBIOME. For these courses, the training set corresponds to the first three weeks and a half, and the test set to the last week of observations. In Figure 2 is possible to visualize a time series plot with the training set and test set for the ENGINF course.

Given the large number of combinations of training sets and models (five courses and three techniques), this section shows the plot of the time series prediction for the three techniques applied to the data of a specific course: ENGINF (Figure 3).

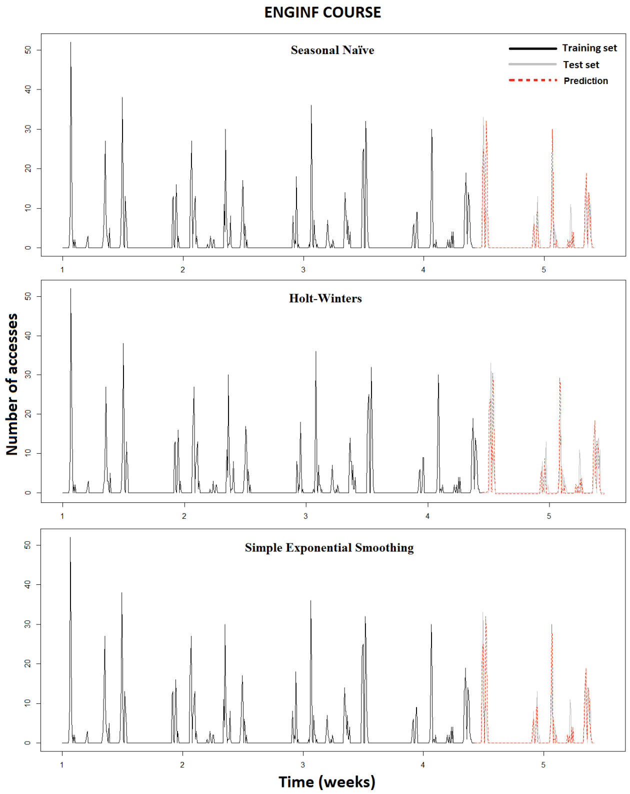

Figure 3: Predictions results for an identified course

To access the techniques, Table 2 summarizes the errors of the three models for the five selected courses. As can be seen, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to evaluate the accuracy of the models and compare them. Specifically, the RMSE will allow analyzing when a model is preferable to other if it is undesirable to have large punctual errors in the number of predicted accesses.

3.3 Evaluation

The results presented show the accuracy of the techniques tested, since all of them offered acceptable results (the predictions molded themselves sufficiently well to the test sets). Accordingly to the models built, all the techniques managed to forecast the presence of students on the given spot, in the tested period, for several different courses (different time series although of similar behavior in trend, seasonality and cyclic aspects). Consequently, emerges the accomplishment of the business success criteria's defined during the business understanding phase: (a) the grounded achievement of a spot on the campus to situate the interactive public display; (b) a basic identification of the profile of students, specifically their university courses; and (c) an accurate estimation of the number of accesses of students of each course on the selected spot at forthcoming periods. Specifically, the first was achieved during the data understanding phase; the second, identification of the course of students, has been shown as accomplished in the data preparation section before the construction of the time series; and the last goal, the accurate estimation of the number of students of each course in the selected spot at forthcoming periods, was proven through the models errors analysis.

More than the results obtained, the forecasting techniques used provided interesting thoughts and considerations. It is possible to confirm that when the time series in analysis have explicit pattern, it should be chosen a technique adapted to it. Simple Exponential Smoothing could not be a better example of that because although presented the better training accuracy, the lack of adaptability to seasonal time series made it provide worst forecasts. Holt-Winters provided also interesting insights. Although it presented worst results of MAE accuracy, its exponential smoothing characteristics aggregated with its adaptability to seasonal data, made it provide better results for the RMSE, which proves its suitability for forecasting systems that are more sensible to high errors. However, from the three techniques, the one that was considered the most appropriate for the envisioned ubiquitous platform was the Seasonal Naïve. The reason for that came from the merge of the results of accuracy with the rationalization regarding the system characteristics' and the variables involved. The forecasting provided by this technique, which only uses the last seasonal period to produce its results, gave the best MAE accuracy and, for the case, that is considered the best approach. Given that it is involved human behavior and the spot is a bar, it is likely that in some moments/days occur peaks of presence. If it were more important to try to be prepared for those peaks than for being arranged for the average case, probably Holt-Winters would be better. However, Seasonal Naïve should be the choice because it provides the system with the ability of doing better forecasts for the average of persons present, which is considered more important than trying to predict peaks of presence.

After the achievement of the grounded spot and of the accurate forecast (and the insights about the techniques to use), the next

Table 2: Error consolidation table (errors presented in number of accesses)

		Seasonal Naive		***Holt-Winters***		***S. E. Smoothing***	
		MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
ENGINF	Training	1,04	3,38	2,15	4,53	0,57	1,59
	Test	1,32	3,73	1,76	3,16	1,65	3,80
LA	Training	0,55	2,04	0,70	1,81	0,39	1,20
	Test	0,67	2,22	0,65	2,22	0,74	2,20
MBINF	Training	1,02	3,00	1,80	3,56	0,70	1,42
	Test	1,28	3,33	2,40	3,84	1,35	3,27
MEI	Training	4,26	11,37	5,88	10,62	2,87	7,01
	Test	3,32	8,35	3,37	5,36	3,36	7,45
MIEBIOME	Training	1,00	3,75	1,66	4,27	0,60	1,54
	Test	1,46	5,40	2,37	3,39	1,58	4,89
	Total Average	1,59	4,66	2,27	4,28	1,38	3,44
	Training Average	1,57	4,71	2,44	4,96	1,03	2,55
	Test Average	**1,61**	**4,61**	**2,11**	**3,59**	**1,74**	**4,32**

step, for future work, is the one that completes the adaptive business intelligence architecture: the development of the optimization module.

4. CONCLUSION

In this work, adaptive business intelligence techniques and the CRISP-DM data mining methodology were used to: 1) identify the basic profile (course/area of study) of a set of students on a university campus based on past wireless accesses to the EDUROAM network, courses schedules and maps with AP locations; 2) select the best spot on the university campus to set a public interactive display taking into account the public spots with higher wireless accesses, and 3) predict future accesses of students of several courses based on their previous Wi-Fi accesses habits. For future work remains to be built an optimization system that will take as inputs the predicted values of the time series model selected and will maximize the adaptation of the contents on the display to the predicted audience.

Understanding person's wants and needs is assuming each day a greater prominence in the market of technological services. By definition, ubiquitous computing discusses technologies that, providing services to users in an environment, are transparent and adapted to them. To this end, seems obvious the need of supporting systems capable not only of suggesting decisions, but also of assuming automatically grounded resolutions. In this context emerge the Adaptive Business Intelligence that aggregates an integrated set of techniques that endow technological platforms of capabilities of learning, acting, reacting and predicting. This work intended to gather together these two worlds - Ubiquitous Computing and Adaptive Business Intelligence - that seam intended to be together, taking advantage of individual strengths to construct mutual advantages.

5. ACKNOWLEDGMENTS

We would like to thank to the Communications Service and to the Department of Informatics of University of Minho for providing the means to perform this research.

6. REFERENCES

[1] Brillinger, D.R. 2000. Time Series : General. *Int. Encyc. Social and Behavioral Sciences*. November (2000).

[2] Cacciapuoti, A.S., Calabrese, F., Caleffi, M., Di Lorenzo, G. and Paura, L. 2013. Human-mobility enabled wireless networks for emergency communications during special events. *Pervasive and Mobile Computing*. 9, 4 (Aug. 2013), 472–483.

[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Rüdiger Wirth 2000. *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

[4] Chatfield, C. and Yar, M. 1988. Holt-Winters Forecasting: Some Practical Issues. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 37, 2 (1988), 129–140.

[5] Chen, S., Li, Y., Ren, W., Jin, D. and Hui, P. 2013. Location prediction for large scale urban vehicular mobility. *International Wireless Communications and Mobile Computing Conference (IWCMC)* (Jul. 2013), 1733–1737.

[6] Forecasting: principles and practice: 2012. *https://www.otexts.org/fpp*.

[7] Kukka, H., Oja, H., Kostakos, V., Gonçalves, J. and Ojala, T. 2013. What makes you click: exploring visual signals to entice interaction on public displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (New York, USA, 2013).

[8] Michalewicz, Z., Schmidt, M., Michalewicz, M. and Chiriac, C. 2006. *Adaptive Business Intelligence*. Springer Berlin Heidelberg.

[9] De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*. 3, (Jan. 2013), 1376.

[10] Ostertagová, E. and Ostertag, O. 2011. The simple exponential smoothing model. *Modelling of Mechanical and Mechatronic systems*. (2011), 380–384.

[11] Vazquez-Prokopec, G.M., Bisanzio, D., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Ramirez-Paredes, J., Halsey, E.S., Kochel, T.J., Scott, T.W. and Kitron, U. 2013. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one*. 8, 4 (Jan. 2013), e58802.

[12] Weiser, M. 1991. The Computer for the 21st Century. *Scientific American*. 265, 3 (Sep. 1991), 94–104.

[13] Wilheim, J. 2013. Mobilidade urbana: um desafio paulistano. *Estudos Avançados*. 27, 79 (2013), 7–26.