UNIVERSITÀ DI TRENTO
Department of
Information Engineering and Computer Science

Polo Ferrari 2 - Via Sommarive, 9 - 38123 Trento (Italy)
http://disi.unitn.it

A survey on students' daily routines and academic performance at the University of Trento

Fausto Giunchiglia, Matteo Busso, Mattia Zeni, Ivano Bison

June 22, 2022

Technical Report # DISI-2001-DS-04

TO BE CITED AS: Fausto Giunchiglia, Matteo Busso, Mattia Zeni, Ivano Bison, A survey on students' daily routines and academic performance at the University of Trento, University of Trento Technical Report. No. DISI-2001-DS-04, (2022)

A survey on students' daily routines and academic performance at the University of Trento

Fausto Giunchiglia[1], Matteo Busso[1], Mattia Zeni[1], and Ivano Bison[1]

University of Trento

Abstract

This technical report describes a dataset which contains *diachronic* data about the everyday life of one hundred fifty-eight university students over a period of four weeks, and also additional *synchronic* data about profile, e.g., demographics, routines, personality. The diachronic data are collected from thirty-four sensors, both hardware and software, associated to around 100+ thousand self-reported annotations. The dataset has been collected based on an ontological representation of the situational context and following various reference standards, e.g., HETUS and the Big Five. The data collection is motivated by the rise of so-called people-centric sensing paradigm, wherein sensors embedded in mobile phones and other wireless devices are used to collect large quantities of continuous data pertaining to the behavior of individuals and social networks. These datasets offer unique opportunities to investigate the diversity and daily routines of university students in a multi-layered perspective.

1 Background & Summary

Academic performance is an important issue for the Italian university system. Empirical evidence shows that academic achievement is a possible cause for students' drop-out and the time required for obtaining a university degree [1]. On one hand, these consequences can affect the chance of finding a (good) job after the degree and, on the other hand, they represent a waste of economic and social resources within a society [2]. Empirical researches have shown how students' time management ability and its translation into time allocation between academic and other daily are important aspects that have an impact on students' performance [3]. This is especially true for higher education institutions in which students have to plan and to organize their study schedule usually without parental support or teacher supervision [4]. However, students' time allocation has received little interest in sociological research [5] and this is partially due to the available data and their reliability. Surveys usually ask to students to report their total amount of time they spend for any given activity, e.g. studying or attending lessons or filling a time diary for a day. These tools suffer from different problems that may affect the quality of time use data. The Smart University project aims at filling this gap in the literature and this document describes the first iteration of the project called SmartUnitn(Two), which was carried out in the University of Trento in late 2016. Although it was designed and run without explicitly following the experiment design methodology, it still is a relevant example of the new type of interdisciplinary experiments.

The final goal of the project (and thus of its iterations) is to understand how the spatio-temporal organization of the students of different degrees affects their academic performances. The analysis will be conducted using personal data about the users, collected from a cross-sectional questionnaire and two type of longitudinal data, deriving from time diaries and sensor data collected from the i-Log app, installed on the students' smartphone.

2 Methods

The overall data collection process lasted six weeks. The process was articulated as a two-stage data collection, as follows:

- The first synchronic data collection, administered through a set of three standard close-ended questionnaire, allowed to collect self-reported general data on material on social practices;
- The second diachronic data collection, administered via a smartphone app, allowed us to observe the students' daily routines.

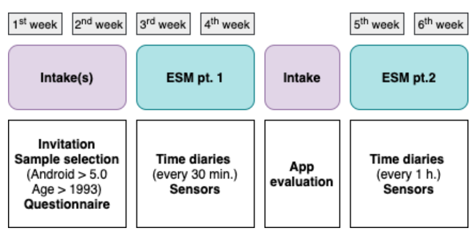

Fig. 1. Schematic representation of the study protocol

As described in Figure 1, the first two weeks were dedicated to the sample recruitment. This was performed by sending two initial questionnaires, i.e., invitation and assessment of habits. The remaining month was entirely dedicated to the data collection through the app installed on the students smartphone. During all the data collection a help-desk was active, and ready to support students in all the problems which were arising.

2.1 Data collection tools

The questionnaires were managed with the LimeSurvey [6] platform. An invitation to participate in the online survey was sent through LimeSurvey to the email address of students enrolled at the various universities. This data collection was based on the use of *Time Diaries*. In the social sciences, time diaries are a classic data collection tool [7], where respondents are asked to indicate three main dimensions of their everyday life: the activities performed, the locations they visit, and the people around them. Time diaries can be administered either as "leave behind diaries", where the respondents fill the data in real time as the day progresses, or as "recall diaries", where respondents must recall their activities for the previous day. Via `iLog app` [8, 9][1], students had the possibility to answer in real time, where questions and answers were managed in a way to follow the HETUS standard[2].

2.2 Sample design

The sample was selected within the entire student population of the University of Trento. All students were sent an invitation to participate in the survey, excluding a priori those who did not have a smartphone compatible with the study (only Android Operating System greater than 5.0) or did not regularly attend classes. After the first contact via email, the online questionnaire was sent to investigate their habits and routines. The last step consisted to send a password to be used to download and install the iLog. This activity resulted into 1042 responses. Out of these responses, we deleted those from students who were born after 1993 (with the goal to limit the dynamics of out-of-school late students) and students who did not actively participate in university life. Of the 860 candidates, 318 students were selected, weighting the sample in such a way that it was proportional to the student population of each department. This last step was taken to avoid the misrepresentation of daily routines due to the different schedules and University sub-communities. As from the title of this paper, this dataset contains data about *only* 158 participants. This reduced size, with respect to the 318 students mentioned above is the consequence of the data cleaning, done during the data preparation step described in the next section, where all the participants with a small participation to the survey were dropped.

2.3 Incentives strategy

The Incentive strategy consisted of paying a fixed salary and giving a series of bonuses. All participants who filled in at least 75% of the questions received €20 every two weeks. Furthermore, they had the opportunity to participate in the draw for three prizes of €100 for the first two weeks of participation and of further three prizes of €150 for the second two weeks. Finally, every day three

[1] [10–14] is a list of publications which describe the use of iLog and of iLog collected data in various experiments. Currently, iLog runs on Android devices; the iOS version is under development. The possibility of collecting both user answers and sensor data makes iLog quite unique (see [15–17] for a list of other tools currently available). This double facility is quite important in that it allows to improve the state of the art in time diaries [18, 19], especially if structured [20].

[2] *Harmonized European Time Use Surveys*: https://ec.europa.eu/eurostat/web/time-use-survey

prizes of €5 were drawn and communicated via email to the winning participants. The goal of the prized was (i) to keep the participants' attention high; (ii) to encourage communication with the help-desk.

3 Data records

The final dataset contains data about *only* 158 participants. This reduced size, with respect to the 318 students mentioned above is the consequence of the data cleaning, done during the data preparation step described in Section 4, where all the participants with a small participation to the survey were dropped. The following paragraphs are aimed at describing the data from the three different sources, namely (i) questionnaire; (ii) time diaries; (iii) sensors.

3.1 The questionnaire

Table 1 shows how the sample is balanced according to the main characteristics, namely gender, age and departments in which the students were enrolled. Furthermore, it shows the range of annotations given from the participants. The psycho-social characteristics of the participants are described in table 2.

Table 1. Descriptive statistics of the participants

	%
Gender	
Female	48.7
Male	51.3
Age	
<22	47.5
22-26	52.5
Departments	
Hard Sciences	37.3
Soft Sciences	33.5
Humanities	29.2
Total	100.0 (N=158)
iLog Obs.	396-932

Concerning personality traits (BFI-10), the average of the scores is between 6.1 and 7.3, with a maximum standard deviation of 2.17 reached in the case of the Neuroticism variable. The range of responses goes from 2 to 10. As for procrastination (IPS), the average is 22.7, with a standard deviation of 6.25, while the range of responses goes from 10 to 40. As for smartphone addiction (SAV-SV), the mean is 27.5, with a standard deviation of 17.6, while the response

Table 2. Descriptive statistics of the psycho-social traits

	mean	sd	range
Agreeableness	6.7	1.74	2-10
Conscientiousness	7.3	1.76	2-10
Extraversion	6.1	1.94	2-10
Neuroticism	6.6	2.17	2-10
Openness	6.9	1.92	2-10
Procrastination	22.7	6.25	10-40
Smartphone Addiction	27.5	17.6	0-93.3
Perceived Stress	43.3	17.4	7.5-85

range is from 0 to 93.3. Finally, as regards perceived stress (PSS), the average is 43.3, with a standard deviation of 17.4, while the range of responses goes from 7.5 to 85.

Other variables collected concerned (i) the routine of daily and extraordinary journeys considering the times and means used; (ii) work routine; (iii) the study and class attendance routine. Therefore, in addition to psycho-social traits, 27 questions were asked, for a total of 78 variables collected, most of which do not have missing data.

3.2 The time diaries

The bar-plots for each of the answers to the time diaries questions are presented below. The total of observations is 168,095, including missing values.
From Figure 2, in 25,084 (24.7% of the total of non-missing answers) cases the students declared that were "Sleeping", answering at the question "What are you doing?". In addition, 17,459 (17.1%) annotations concerned "Studying", while 9,156 (9.0%) annotations were "Eating", followed by all the others. The answers to the question "How are you moving?" are much less, as the question appeared only if the participant claimed to be "En route". Most of these were "By foot" (3.169, i.e. 36.7%). As for "Where are you?", 53+ thousands (more than 66%) of the annotations concerned a home or a home of parents or friends, while the second most used category of answers concerned university places , especially "Classroom" with 9.737 (12.2%) annotations. In general, students found themselves "alone" most of the time, with 41,269 (46.8%) annotations or with friends in 16.7% of cases. Finally, the "mood" was more positive for most of the annotations (52,527, i.e. 59.5%), while it was rarely really negative, only in 1.7% of cases.

3.3 The sensor data

The sensor data collected are rather rich and diversified. We are not aware of any other dataset with similar properties. Furthermore some of the sensor selected are somewhat unusual. The data from the sensors can be divided into:

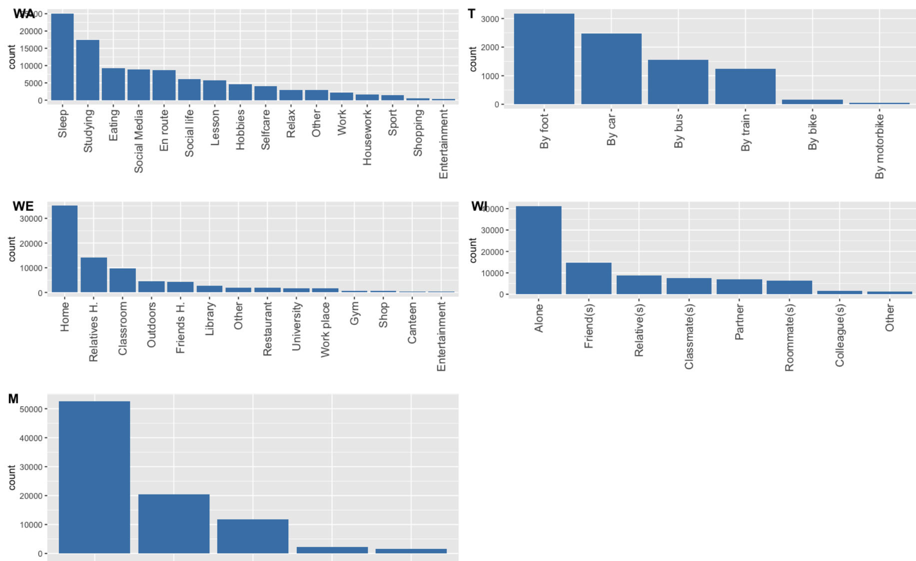

Fig. 2. Amount of answers for each questions: "What are you doing"(WA), "How are you moving?" (T), "Where are you?" (WE), "With whom are you?"(WI), and "What is your mood?" (M).

– Hardware (HW) sensors, that in our case corresponded to the two version of the location, namely GPS(rounded down) and Point Of Interest (POI) reported in Table 3;
– Software (SW) sensors, by which we mean all the SW events that can be collected from the Operating system and SW, for instance the Wifi the HW is connected and so on. The complete list of SW sensors is reported in Table 4.

Table 3. HW sensors.

No	HW Sensor	N. Obs.	Estimated Frequency	U.M.
1	Position	4.144.214	Once every minute	degrees, minutes, seconds
2	POI	1.583.389	Once every 5 minutes	Unitless

In these two sensor tables, the frequency by which the sensors are captured is reported, according to the following conventions: *on change* means that the value of the sensor is recorded only when the current value is changed (along with a timestamp of when it happened), *up to X samples per second* means that for each second the value of the sensor will be stored up to maximum of X times (these values are estimated), and *once every Y* means that the values of a sensor is recorded once the time Y has passed (these values are estimated). The meaning

Table 4. SW sensors.

No	SW Sensor	N. Obs.	Estimated Frequency	U.M.
3	Audio mode [Silent/Normal]	2.042.901	On change	Unitless
4	Battery Charge [ON/OFF]	42.664	On change	0/1
5	Battery Level	26.934	On change	%
6	Doze Modality [ON/OFF]	11.914	On change	0/1
7	Flight Mode [ON/OFF]	2.567	On change	0/1
8	Headset plugged in [ON/OFF]	171.677	On change	0/1
9	Music Playback (no track information) [ON/OFF]	92.510	On change	0/1
10	Notifications received	3.224.577	On change	Unitless
11	Proximity	11.405.724	On change	0/1
12	Running Application	35.184.768	Once every 5 seconds	Unitless
13	Screen Status [ON/OFF]	1.252.576	On change	0/1
14	WIFI Network Connected to	747.366	On change	Unitless
15	WIFI Networks Available	2.859.187	Once every minute	Unitless

of the sensors is described in the codebook that can be downloaded from the dataset Catalog.

4 Technical validation

The data preparation consisted of two main activities:

1. *Data cleaning,* with the goal of consolidating data and bringing them up to the desired quality
2. *Anonymization,* aiming at removing any type of personal information from the data set

In the first phase sensor data were only modified and simplified in their internal format to make it easier to manipulate them. The data cleaning of the online questionnaire consisted of the normalization of all the value labels, where all the relevant personal information was deleted (e.g., email) or substituted with an unique identifier (personal names). The details of the anonymization are reported in the next subsection. As to the time diaries, the total amount of recorded annotations were 139.239. Of these, 107.850 were collected in the first two weeks and 31.389 in the second two weeks. However, not all the students have completed the data collection. Of the 307 students contacted, 273 completed the installation procedure correctly, while only 237 provided answers for more than two days. For this reason we have decided to keep only the participants who:

- completed the time diaries for at least 14 days (both for the firsts and the seconds weeks)
- provided at least 300 answers for the firsts weeks
- provided at least 100 answers for the seconds weeks

The final number of valid subjects for the first two weeks was 150 out of which 88 were also selected for the second two weeks. In the end, 158 participants have been included in the dataset, where the extra 8 participants were identified by slightly relaxing the requirements for the first two weeks, given their very good performance in the second two weeks.

The anonymization procedure dealt with *all* possible sources of possible identification, namely: (i) Personal Data Anonymization, (ii) Network Anonymization and (iii) GPS Anonymization. Notice however that this is not enough, in the sense that this dataset, like any other dataset of the same kind, would lead to re-identification if cross-checked with other datasets. Because of this, this dataset is not published open in the Web. This dataset can only be used in isolation and under certain precise conditions (see below the section on distribution).

For what regards the *Personal Data Anonymization*, all personal information, i.e., *e-mail address*, *home address*, *name and surname*, has been removed from each of the three types of datasets (online questionnaire, time diaries and sensors), still making sure that the same unique identifier would be assigned to the same person across all three datasets.

For what concerns *Network Anonymization*, there are three possible sources of re-identification, namely: (i) *WiFi-event*, that shows the WiFi network the smartphone is connected to; (ii) *cellular-network*, that shows the roaming network; and (iii) *WiFi-networks-event*, that shows the WiFi networks that are available in the environment. For each of these sensor files, the relevant columns have been anonymized via the use of unique identifiers. A hash function was applied to the WiFi network name, with a function that cannot be reversed (the SHA-256 cryptographic function is used to perform the hashing).

For what concerns *GPS Anonymization*, the main problem is that the position of a person, in particular if joined with the specific time and day, leads very easily to re-identification, in particular when a person is in places which are not too crowded (e.g., outside cities) or when collected for a long period of time. The only solution is to make the spatio-temporal information more ambiguous. There are many ways of doing this, all having different consequences of the usability of the dataset for research. The GPS information of this data set has been anonymized in two different ways

1. *Round Down* Here the idea it that precision is deliberately truncated from the location sensor so that it becomes anonymous but in a way to be still useful for certain scientific purposes. Furthermore, the dates associated to each GPS point are truncated;
2. *Point of Interest (POI).* Here the idea is to collect *only* those points where the user has spent more than a certain amount of time. In this dataset, if latitude and longitude do not change for one minute then a POI tag is added to the stream. For each POI the elapsed time in seconds is also added. GPS longitude and latitude readings are removed. The POI is selected to identify a general location (suburb, city, region) and the closest relevant places (bar, restaurant, lake, etc).

These procedures have produced two sets of datasets, that we call *RoundDown* and *POI*, each containing all the other sensors. For privacy reasons, only one of the two datasets can be downloaded by the same research institution (the union of the two would easily lead to re-identification).

5 Usage notes

The dataset is available in different formats, depending on the type of data:

- Questionnaire and time diaries: they are available in the different formats supported by the main software, namely STATA (.dta), SPSS (.sav), R (.rds) and EXCEL (.xlsx)
- Sensors and time diaries: they are available in .csv format and in PARQUET.

The main entry point documentation for this dataset is the web-page:

http://livepeople.datascientia.eu/workspace/smartunitntwo

The website contains links for further information regarding:

1. The tools used, including the link to the questionnaire and the iLog documentation;
2. Description of the labels and the values of the variables
3. Descriptive and summary statistics of the collected data

Additionally, the website contains a sample of the dataset and links to all articles that have been published through the dataset. Further documentation and metadata will be provided in the project catalog.

Because of the type of data, to be fully compliant with `GDPR`, in order to have access to any of the datasets described in this paper, a licence must be signed. The details of how to enable this will be provided after contacting us at the following email `datadistribution.knowdive@unitn.it`. Some relevant licensing conditions are: (i) the datasets may only be used for research purposes; (ii) redistribution of the datasets is forbidden; (iii) the datasets cannot be made public (e.g., on a website) or given to a third part.

6 Code availability

Any scripts for the data preparation or for the analysis of the dataset will be shared on the catalog.

References

1. Moris Triventi and Paolo Trivellato. Participation, performance and inequality in italian higher education in the 20th century. *Higher Education*, 57(6):681–702, 2009.

2. A Schizzerotto and F Denti. Perduti e in ritardo. l'esperienza dell'abbandono e dell'irregolarità degli studi in cinque leve di immatricolati all'ateneo di milano-bicocca. *Rapporto di ricerca*, 2005.
3. Sahebe Yousefi, Zahra Hrsej, Zahra Jannat Alipour, Mahmoud Afrrouz, and Nasrin Navabi. The relationship between time management skills and academic achievement in nursing and public health students. *Medical Education Journal*, 5(2):29–40, 2017.
4. Shazia Nasrullah_PhD and Muhammad Saqib Khan_PhD. The impact of time management on the students' academic achievements. 2015.
5. Alain Fernex, Laurent Lima, and Erica De Vries. Exploring time allocation for academic activities by university students in france. *Higher Education*, 69(3):399–420, 2015.
6. LimeSurvey Project Team / Carsten Schmitz. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany, 2012.
7. John P Robinson. The time-diary method. In *Time use research in the social sciences*, pages 47–89. Springer, 2002.
8. Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. Multi-device activity logging. In *Proceedings of ACM - UbiComp*, 2014.
9. Mattia Zeni, Ivano Bison, Britta Gauckler, Fernando Reis, and Fausto Giunchiglia. Improving time use measurement with personal big collection - the experience of the european big data hackathon 2019. *Journal of Official Statistics*, 2020.
10. Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 299–302, 2014.
11. Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. Mobile social media and academic performance. In *International conference on social informatics*, pages 3–13. Springer, Cham, 2017.
12. Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Personal context modelling and annotation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 117–122. IEEE, 2017.
13. Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Human-like context sensing for robot surveillance. *International Journal of Semantic Computing*, 12(01):129–148, 2017.
14. Fausto Giunchiglia, Mattia Zeni, and Enrico Big. Personal context recognition via reliable human-machine collaboration. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 379–384. IEEE, 2018.
15. Jason D Runyan, Timothy A Steenbergh, Charles Bainbridge, Douglas A Daugherty, Lorne Oke, and Brian N Fry. A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. *PloS one*, 8(8):e71325, 2013.
16. Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14, 2014.
17. Frauke Kreuter, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, 38(5):533–549, 2020.

18. Fausto Giunchiglia, Mattia Zeni, Enrico Bignotti, and Wanyi Zhang. Assessing annotation consistency in the wild. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 561–566. IEEE, 2018.
19. Pitirim Aleksandrovich Sorokin and Clarence Quinn Berger. *Time-budgets of human behavior*, volume 2. Harvard University Press, 1939.
20. Mattias Hellgren. Extracting more knowledge from time diaries? *Social Indicators Research*, 119(3):1517–1534, 2014.

7 Author contributions statement

The order of names is by contribution of the Institution and, inside each Institution, by contribution of the individuals. As such, the order of names does not necessarily reflect the importance of the contribution of the single individuals. The roles of the authors, presented by by their initials, is as follows:

- *Study management*: F.G., I.B.;
- *Study design*: F.G., I.B.;
- *Technical support*: M.B., M.Z.;
- *Data Collection*: M.B., M.Z.;
- *Data Preparation and correction*: I.B., M.B., M.Z.

8 Acknowledgements

This research has received funding from the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us", grant agreement No 823783. This research has deeply benefitted from the huge number of discussions that we have had with all the people working in the WeNet project.

9 Competing Interest

All authors declare no competing interests during the data collection, preparation and analysis of this dataset.