Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 10835

50x2030 Data-Smart Agriculture Series

Making Time Count

A Machine Learning Approach to Predict Time Use in Low-Income Countries from Physical Activity Tracking Data

Joris Mulder
Seyit Hocuk
Talip Kilic
Alberto Zezza
Pradeep Kumar

WORLD BANK GROUP
Development Economics
Development Data Group
June 2024

Policy Research Working Paper 10835

Abstract

Understanding men's and women's time use is a key factor in addressing issues and formulating policies related to division of labor, domestic work, and related gender disparities. However, obtaining data on individuals' time use can be difficult and costly in the context of household surveys. Leveraging unique survey data collected in rural Malawi, this study investigates the possibility of predicting men's and women's time allocation to an extensive set of activities, using sensor signal data captured by accelerometers. Using machine learning techniques, the study builds a supervised classification model that is trained on the accelerometer data and a random subset of the time use survey data to predict individuals' time allocation to 12 broad activity groups. The model can correctly classify each performed activity in 76 percent of the cases. The analysis shows that with 40 percent of the training data, this method can achieve 90 percent of the maximum level of predictive accuracy reached in the analysis. The findings prove the feasibility of this methodology and offer insights for enhancing both survey and accelerometer data collection processes to build better models. Using the method can improve the quality of costly and difficult to obtain time use surveys with cheaper, yet accurate, modeled estimates, obtained by combining objective data from wearable devices with time use data collected on smaller samples.

This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at tkilic@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Making Time Count

A Machine Learning Approach to Predict Time Use In Low-Income Countries from Physical Activity Tracking Data

Joris Mulder[†], Seyit Hocuk[+], Talip Kilic*[‡], Alberto Zezza[‡], and Pradeep Kumar[†1]

JEL Codes: C63, C83, J22.

Keywords: time use, accelerometer data, activity recognition, predictive modeling, machine learning, classification models.

[1] * Corresponding author: tkilic@worldbank.org. ‡ Development Data Group, World Bank; † Centerdata, the Netherlands; + Fabory, The Netherlands. The order in which the author names appear has been randomized. This paper was produced with the financial support from the World Bank Living Standards Measurement Study (LSMS), the World Bank Research Support Budget (RSB), and the 50x2030 Initiative to Close the Agricultural Data Gap, which is a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030.

1 Introduction

"Time is Money" wrote Benjamin Franklin nearly 300 years ago (Franklin, 1748). Indeed, for welfare economists a fundamental argument in utility and welfare models, alongside command over economic resources, involves the allocation of time, most notably the quantity of time devoted to leisure and other non-work activities (Wales and Woodland, 1977; Killingsworth and Heckman, 1986; Pencavel, 1986; Fafchamps, Kebede, and Quisumbing, 2009). Labor economists for their part are increasingly focusing on understanding and valuing that portion of work that does not result in market transactions, such as work spent on domestic chores as well as on household own production (Gaddis et al., 2023; Folbre, 2021; ILO, 2013; Beneria, 1999). As part of the Agenda 2030, Sustainable Development Goal (SDG) 5 on girls' and women's empowerment includes a target on valuing domestic and unpaid care work, to be monitored via indicator 5.4.1, the proportion of time spent on unpaid domestic and care work, by gender, age, and location.

Global patterns in men's and women's time use demonstrate the existence, particularly in low-income countries, of significant disparities in the time that men and women devote to activities such as unpaid domestic and care work and activities that are more closely tied to income generation. On average, women spend at least twice as much time as men in unpaid work, and often have a higher total work burden than men, when both paid and unpaid work are considered (UNDESA, 2010). These disparities impose several constraints on women's well-being and livelihoods. Notably, women who choose to enter the labor market must contend with serious trade-offs in terms of their own personal needs (e.g., employment, education, leisure, personal care) and those of their families. Children's diets and nutrition-related health behaviors are heavily dependent on mothers' time allocation to a range of economic and nutrition- and health-enhancing activities (Black et al., 2013).

Efforts to formulate policies to address issues relating to income-generating activities, time allocation of domestic work, and gender disparities require accurate data on men's and women's time use. The revised statistical definition of 'work' which includes unpaid care and domestic activities (ILO, 2013), and Sustainable Development Goal Target 5.4 to "recognize and value unpaid care and domestic work through the provision of public services, infrastructure and social protection policies" have provided additional momentum for investments in time use data collection.

These analytical objectives, and the design of effective policies and programs for addressing gender disparities in both paid and unpaid work that they aim to inform, rely on the accurate measurement of men's and women's time use. The existing and persisting weaknesses in individual-level data and statistics on time use (Chatzitheochari et al., 2017; Harvey and Taylor, 2000; Juster et al., 2003; Kan and Pudney, 2008; Kes and Swaminathan, 2006) are a major bottleneck hindering progress towards these policy research goals. Time use surveys or time use modules as part of multi-topic household surveys are key sources of the required data. It is most common for these surveys and modules to elicit respondents' recollection of their activities during the reference period, typically the previous 24 hours leading up to the interview (Fisher, 2015).[2]

The shortcomings in terms of cross-country availability, comparability, and quality of individual-level data on time use can in part be traced back to underinvestment in rigorous methodological research for the development of scalable and accurate survey methods (Daum et al., 2019; Kilic et al., 2024). Research has drawn attention to the weaknesses in the availability and quality of individual-level survey data and statistics on time use measurement (Chatzitheochari et al., 2017; Harvey and Taylor, 2000; Juster et al., 2003; Kan and Pudney, 2008; Kes and Swaminathan, 2006). Studies have emphasized the importance of developing improved survey methods for time use (Bauman et al., 2017; Buvinic and King, 2018; Charmes, 2015), particularly in low-income contexts with high prevalence of illiteracy and unfamiliarity with clock-oriented time (Seymour et al., 2020).

[2] Time use diaries can also be self-administered, and respondents can use a paper diary or an electronic device (e.g., tablet or smartphone) and are instructed to use it to record all their activities during the reference period.

Against this background, this paper reports on a proof-of-concept study that aims to evaluate whether machine learning and accelerometer technology can be leveraged to accurately predict time use patterns from physical activity sensor output collected in household surveys in low-income contexts. We derive our insights from a distinctive survey of agricultural families in rural Malawi. This survey combined twice-repeated 24-hour recall-based time use data collection with concurrent sensor-driven physical activity monitoring. Our paper sits at the intersection of the research streams on time use measurement and physical activity tracking in household surveys, and the main contributions are threefold. First, we build and validate a machine learning model within our sample, by pairing the high-frequency physical activity tracking data with time use survey data for the same individuals. Specifically, we train the model for a subset of the individuals and infer the time use in the rest of the sample and compare the predicted versus observed time use outcomes. If successful and cross-culturally validated, this framework offers survey practitioners the possibility of increasing accuracy in time use data while managing respondent burden by combining data from low-cost, waist-worn, inconspicuous accelerometers and time use data collection from a subsample of survey respondents. Machine learning-based imputed estimates of time allocation to specific activities can then be obtained for the part of the sample that is not part of survey data collection on time use.

Second, by combining time use and sensor data, we identify issues in recall-based time use data collection that provide motivation for improving the activity lists used in surveys and assessing the accuracy of 24-hour recall. This can potentially motivate pairing real-time physical activity tracking with real-time self-administered time use diaries, inspired by the successful experience with smartphone-based time use data collection (henceforth referred to as mobile time diaries) elsewhere in Africa (Daum et al., 2019; Kilic et al., 2024).

Third, the paper brings in evidence from a rural, low-income context regarding the feasibility and utility of physical activity tracking data collection for time use measurement – a topic that has not been the subject of previous research – based on a significantly larger sample of human subjects vis-à-vis earlier work in low-income countries. While accelerometers[3] are widely used to document physical activity levels for specific population groups (see Troiano et al., 2008 and Colley et al., 2011, and the studies cited therein), the vast majority of accelerometer data collection campaigns has been implemented in high-income contexts. The uses of accelerometers in low-income contexts, and specifically as part of socioeconomic surveys and development research, have been rare events, except for a handful of small-scale studies (Benefice et al., 2001; Assah et al., 2011; Christensen et al., 2012; Zanello et al., 2017; Picchioni et al., 2020; Srinivasan et al., 2020).

The paper is organized as follows. Section 2 describes the data collection steps and provides descriptive statistics of the data. Section 3 presents the methodology of data preprocessing, and the machine learning models. Section 4 discusses the results, focusing on model outcomes, model and data validation, and insights into activity patterns of various population groups. Section 5 applies the trained model on actual data to obtain the time allocation of Malawian people. Section 6 concludes and draws implications for future work.

2 Data

2.1 The datasets

The data used in this paper were collected in the Ntcheu and Zomba districts of Malawi in 2017 as part of a larger methodological survey experiment on measurement of agricultural labor, managed by the World Bank Living Standards Measurement Study (LSMS). This was in collaboration with the World Bank's Gender Group, Development Research Group, the Active Living Research (ALR) team at the

[3] See Chen et al. (2012) for a scientific review of devices and technologies from the point of view of physical activity researchers.

University of California, San Diego (UCSD), and Wadonda Consult. The project spanned the period of November 2016 to June 2017 and yielded a unique dataset which includes a measure of physical activity of 415 adults (>15 years old) sampled from 215 rural households across 20 enumeration areas in the two districts (Pratt et al., 2020; Friedman et al., 2023).

The physical activity tracking data collection was spread over the period of February-April 2017 and was piggybacked onto the weekly visits that were paid to these 215 households for collecting weekly individual-disaggregated data on plot-level household agricultural labor inputs throughout the 2016/17 rainy season. The field staff were trained intensively on the activity tracking component prior to the start of the fieldwork. This training was led by a team of technical specialists on physical activity tracking from the Active Living Research laboratory at UCSD, who also monitored the incoming physical activity data daily and provided feedback to the field staff on the quality of the incoming data.

During fieldwork, each adult household member that was engaged in agriculture wore a research-grade ActiGraph physical activity tracker for a two-week period during non-sleeping hours of each day. The enumerators visited the sampled households three times during the 17 days when the accelerometers were with the participants, typically days 1 (deployment), 9 (check in) and 17 (pick up). On day 1, the enumerators explained the study objectives to the participants, obtained their consent and demonstrated how to wear the accelerometers (with a belt around their waist). Day 9 was a check-in visit, during which data were downloaded and enumerators checked on battery status and for any unusual patterns in the data (which could signal that the devices were not worn correctly). On day 17 the remaining data were downloaded, and the accelerometers returned for reassignment to subsequent participants.

The participants in the study wore an accelerometer around the waist. Physical activity can be detected using accelerometers worn on various parts of the body, such as waist (hip), ankle, or wrist (Doherty et al., 2017; Menai et al., 2017; Willetts et al., 2018). Waist and wrist are the most frequently used body locations, and both have demonstrated to be valid locations for measuring and classifying physical activity intensity and sedentary time (Romazini et al., 2012; Hildebrand et al., 2014; Chandler et al., 2015). There are a few studies about how the physical activity detection using wrist and waist-worn accelerometers correlate or differ. Kamada et al. (2016) and Dieu et al. (2016) report moderate-to-high correlations between acceleration values from either wrist or waist. Hildebrand et al. (2014) find that wrist-worn accelerometers produce higher output as compared to waist-worn accelerometers. Shiroma et al. (2016) conclude that the waist-worn and the wrist-worn monitors may be employed for different behaviors during the day. Waist-worn devices may measure the whole-body movements while wrist-worn devices measure most of the whole-body movements in addition to the movements being specific to wrist or upper body. Zhang et al. (2012) used a wrist-worn GENEA with a waist-worn GENEA to compare classification accuracy for physical activity into walking, running, household, or sedentary activities. They report that the results from both devices are comparable. It becomes clear from these studies that while wrist worn accelerometers may offer higher accuracy for some activities, waist-worn devices offer a suitable alternative at low cost and with low profile to collect physical activity data in low-income contexts.

Complementing the activity tracking data, each participant also responded to a 24-hour recall time use module at the second and third visit (Days 9 and 17, respectively), which was modeled after the time use module of the Women's Empowerment in Agriculture Index (WEIA). The module intended to capture each respondent's main and secondary activities in 15-minute intervals during the reference day, relying on a list of 26 activities (Appendix A). One of the activities 'Plan the household's finances or bills' was never reported. Only 25 activities are thus used during the analyses. The time use activities broadly describe how Malawian people allocate time in their daily lives, such as 'Farming' and 'Cleaning house'. The time use activities can therefore be considered as broad group activities. **Figure 1** shows the relative subdivision of the 25 activities that is derived from the number of reported minutes in the time use survey.

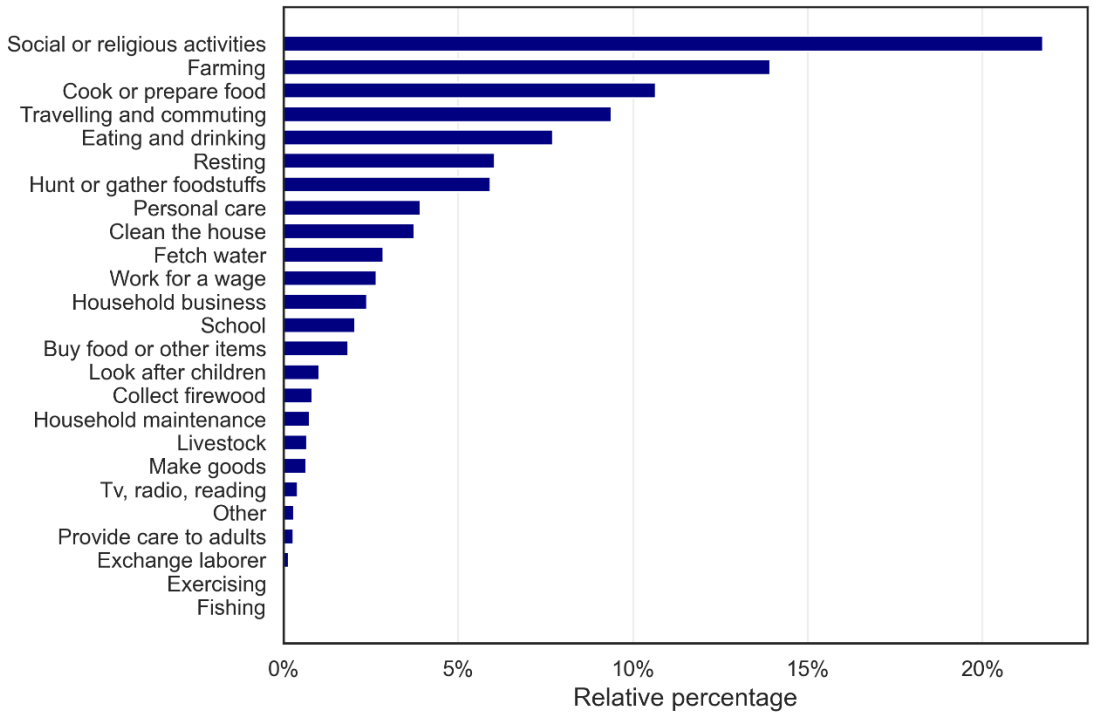

Figure 1. Percentage-wise distribution by the reported minutes of the time use activities from 415 individuals.

Furthermore, the interviewers also measured the participants' bodyweight on the second and third visits, while the data on height as well as household agricultural production, consumption, and expenditures were collected during the endline survey (May-June 2017). This information was complemented with individual-level information on labor market outcomes, collected at baseline, mid-line, and endline interviews.

Overall, the study yielded an average of 13 days of valid activity tracking data across the participants (out of 14 possible full days of physical activity data collection), with an average daily wear time of 14 hours per person. As such, the degree of compliance with accelerometer protocols was significantly higher than studies in higher-income contexts (Pratt et al., 2020). The physical activity tracking data were captured on the sub-second level but exported at the minute level (by taking the mean) according to accepted international standards and protocols (Kerr et al., 2013). The physical activity sensor output of each individual was time-stamped and recorded for each minute of the total wearing time of each day. The time-stamped physical activity tracking data were in turn matched with the recall-based and time-stamped time use data collected for two reference days for each individual. The 15-minute level data of the time use module were disaggregated to one-minute level intervals to ensure matching with the physical activity data. Next to acceleration along three dimensions, the device also infers the number of steps taken per minute.

2.2 Potential data limitations

There are several features of the data that may limit building a model that achieves high prediction accuracy.

The physical activity sensor signal data were collected at minute level, which limits the possibility of incorporating an extensive range of time and frequency domain features of the signal (e.g., power of signal, Fourier transform, and other statistical features). Such features can potentially enhance the predictive power of a model so that hard to discern activities can be recognized with higher accuracy. Other accelerometer studies (Zhang et al., 2012; Mannini et al., 2013; Ellis et al., 2016; Willetts et al., 2017) collected the data at sub-second level and used a wide range of time and frequency domain

features to train their machine learning models. High frequency data may aid in predicting some activities more reliably. The study by Staudenmayer et al. (2009) showed that high frequency data can help in better predicting repetitive activities like cycling in free living conditions.

Regarding the time use data, the interviewers were trained to elicit a narrative of each participant's reference day, establishing first when the participant woke up and went to bed, and then recording the type and timing of the activities in between, in accordance with the respondent's recollection. The time use module allowed for the specification of both main and secondary activity for each 15-minute interval. This feature of the module aimed at capturing simultaneous activities.

Having said that, recall-based time use reports are prone to potential errors, such as bad recollection of the performed activity or attributing the activity to the wrong time interval (e.g., Bell et al., 2019). This way of labeling the data also has the caveat that simultaneous activities can get mixed (Seymour et al., 2020). Especially, for short duration activities there is a possibility that more than one activity can be performed within the 15-minute time window. Hence, overlap in the activities can lead to mislabeling the physical activity sensor signal data but also to rounding errors of the time frames in which different activities are performed.

Furthermore, the activity list for the time use module does not precisely identify tasks or activity levels but are meant to provide a broader classification of how people spend their time. For example, "Work for wage" could mean any number of different activities like working in the field (active work), in a school (less active), or in an office (near sedentary). The same applies for 'Traveling and commuting'. Therefore, predicting some time use activities using physical activity sensor signal data can be more difficult than other, more exclusive activities like 'Fetching water', 'Watching TV', or 'Resting'.

All these limitations can affect the performance of machine learning models. This study therefore also aims to assess the extent to which broadly grouped activities can be predicted using sensor signal data from wearable devices in combination with background information like age, gender, and the time of day.

3 Methods

To achieve the goal of predicting time use activities using real-time physical activity sensor data, we leverage the power of machine learning techniques. In the past 'vector magnitude' methods have been employed to objectively gauge the intensity of physical activity from accelerometer data in large datasets (e.g., Mannini et al., 2013; Doherty et al., 2017; Menai et al., 2018; Kapteyn et al., 2018). More recent studies also employ machine learning methods in large scale accelerometer datasets to objectively measure specific activities (Willetts et al., 2018; Sundararajan et al., 2021; Nunavath et al, 2021; Pfeiffer et al., 2022), demonstrating the broad applicability of such methods.

Some of these studies effectively used machine learning techniques in similar contexts. For example, Willets et al. (2018) conducted an extensive study on 24-hour level physical activity on 96,220 UK Biobank participants. Utilizing a Random Forest machine learning algorithm trained on data from 132 adults, they were able to classify six distinct activities based on accelerometer data, which are bicycling, sitting/standing, walking, driving, sleeping, and mixed activities. Impressively, their model achieved an average prediction accuracy of 87 percent. Furthermore, their research highlighted variations in activity patterns on an hourly, daily, and seasonal basis among adults aged 45 to 79. In a more recent study, Sundararajan et al. (2021) employed a Random Forest model to identify sleep patterns using accelerometer data. Their model differentiated between sleep and wake states with a 74 percent score on unseen data. Moreover, by integrating their sleep detection algorithms with day-time data, they adeptly utilized machine learning to ascertain whether the accelerometer was being worn, achieving a score exceeding 93 percent. Both studies underscore the precision with which machine learning models can predict physical activities.

The main distinction lies in the fact that these studies primarily concentrated on specific, more easily identifiable activities, such as walking, running, driving, and sleeping. In contrast, our work, focuses on non-specific, broadly defined time allocation activities, like farming, taking care of children, working for wage, and collecting foodstuff.

3.1 Dataset merging

Building a supervised machine learning model requires data on observations and relevant predictor variables. Data containing known outcomes, serving as dependent variables, are essential for model building and evaluation. We utilize the reported time use activities collected through the time use survey module as our dependent variable.

The first step in preparing the data for machine learning modeling involved merging the time use survey data with the physical activity sensor data. The sensor-based physical activity tracking study spanned over two consecutive weeks. During this period, participants were surveyed twice with 24-hour recall-based surveys to gain insight into their time use activities. The amalgamation of the datasets took place on the participant level for these two days. The aggregated dataset was further augmented with demographic information of the participants as well as meta information derived from the timestamp.

Harmonizing the data registration time intervals was crucial. The physical activity sensor data were recorded at one-minute intervals, whereas the time-use data were recorded at 15-minute increments. To synchronize the datasets at a consistent granularity, the time-use data were restructured to align with the one-minute interval of the sensor data.

3.2 Dataset cleaning

The combined dataset underwent cleaning, which involved discarding data points with missing or unclear information on time use activities. Subsequently, data points indicating the accelerometer was not being worn were also removed. The wearable device registers inactivity when detached from the body, such as during bathing, and the corresponding time stamps registered by the device as non-wear were marked and discarded accordingly. One additional consideration was that approximately 10 percent of the entries had reported secondary activities simultaneous with the primary activities. From a modeling standpoint, these dual activities introduce more uncertainty instead of clarity. Therefore, such entries were removed from the analysis.

Lastly, we scrutinized the minute level data to identify extreme outliers within 15-minute level epochs. The data-points deviating more than two-sigma from the sample distribution were discarded. This means that all data points diverging from the general behavior pattern of a specific activity within the 15-minute epoch were excluded from the dataset. The discarded outliers accounted for 2.6 percent of the data. The remaining data were considered suitable for the scope of the current study. The cleaned dataset comprised 388 unique participants, indicating, that 27 participants were removed during the cleaning process from the original sample size of 415 participants.

3.3 Feature engineering

The raw data comprises four base variables: accelerations along three axes and the step count. Additional features were engineered from the extracted base variables of the triaxial accelerometer. In the realm of data science, a feature is a characteristic of data. These are used to enhance data analysis. In addition to the base and generated features, we incorporated extra features based on background information and on the timestamp of the time-use activities, representing the day of the week (weekday or weekend day) and segment of the day (am or pm).

We conducted a correlation analysis on the feature engineered data to avoid unnecessary growth of the data space. Correlated variables can add more weight to a variable without adding more value to a model. All highly correlated variables were identified and excluded from the dataset. A Pearson's correlation (or anticorrelation) with a threshold *r*-value of 0.95 was adopted.

The result is a set of 37 features, out of which 28 are derived features generated from the four base variables. These engineered features include the vector magnitude along the three axes, the 15-minute epoch means and standard deviations along the three axes, vector magnitude, and step count, the angles of inclination calculated from the acceleration components, like approximations of roll and pitch, and the direct products of the four base variables.

The complete list of features is summarized in **Table 1**. Appendix B provides more details on each of the features.

Table 1
List of features present in the dataset.

Base features (4)	Background features (5)	Derived features (28)
X acceleration	Age	Vector magnitude
Y acceleration	Gender	Angle variables (7)
Z acceleration	Weight	Direct products (10)
Step count	Day of week	Epoch means (5)
	Time of day	Epoch standard deviations (5)

3.4 Model building and evaluation

The distribution of data points for the time use activities is skewed with respect to the number of reported minutes. Among the time use activities, 12 are reported more frequently, comprising over 90 percent of the data (see also **Figure 1**). The remaining activities are infrequently reported; thus, their omission would not compromise the predictive model performance in ascertaining correct time use activities. Our analyses are therefore performed on the twelve major activities of the dataset only.

Nevertheless, within this subset of major activities, some are still more prevalent than others, resulting in an imbalanced dataset. Training a machine learning model on an imbalanced dataset can lead to biased results. To address this issue, we adopted a strategy of random down-sampling of the more commonly reported activities, ensuring uniform reporting instances (in terms of minutes) across all activities. The resulting dataset is then divided into a train set (80 percent) and a test set (20 percent). The train set is used to build the model, while the test set is used to evaluate its performance of the model. All data values are normalized prior to train-test splitting using a min-max scaling to give equal weights to all features.

Our analyses leverage two machine learning classifiers: Gradient Boosting and Random Forest. For Gradient Boosting, we utilize the XGBoost implementation in Python.[4] XGBoost is a highly effective and widely employed machine learning algorithm used in many challenging machine learning problems. By selecting a random subsample of the main dataset without replacement for each iteration, we effectively enforce the Stochastic Gradient Boosting (SGB) method. For Random Forest, we apply the scikit-learn implementation (Breiman, 2001). The Random Forest classifier is a popular machine learning model used in many scientific studies. The algorithm applies the general technique of bootstrap aggregating to tree learners, also known as bagging (Breiman, 1996).

Model performance hinges on multiple metrics. In this work, we primarily focus on prediction accuracy and the F_1-score. All scores are based on the unseen test data. The F_1-score represents the harmonic mean of precision and recall, optimizing for true positives of the prediction while minimizing false positives and false negatives. Since the evaluation metrics are provided on the activity level, we adopt the strict macro average class score as the ultimate measure of model performance.

[4] https://xgboost.readthedocs.io/en/latest/python/python_api.html

4 Results

4.1 Predicting time use activities

We present the results for the two utilized predictive models, SGB and Random Forest. We find that our SGB model results in a macro average F_1-score of 76 percent. Our Random Forest model results in an F_1-score of 68 percent. As a basis for comparison, a random guess would result in a prediction accuracy and an F_1-score of merely 8.33 percent since there are 12 balanced classes.

From our findings, we can conclusively assert the following:

1. Machine learning effectively utilizes accelerometer data to forecast time use.
2. The SGB model outperforms the Random Forest model.

Consequently, we have selected the SGB model as our best performing model, and all subsequent model results presented will pertain exclusively to the SGB model. Detailed prediction scores for the SGB machine learning model can be found in **Table 2**.

Table 2
SGB model prediction scores and test set characteristics for the twelve main activities.

Activity	Set size	Participants	Precision	Recall	F_1-score
1. Social or religious activities and hobbies	2036	333	68 %	54 %	61 %
2. Farming	2036	199	75 %	77 %	76 %
3. Cook or prepare food or drinks	2036	239	56 %	69 %	62 %
4. Travelling and commuting	2036	316	80 %	55 %	65 %
5. Eating and drinking	2036	380	68 %	55 %	61 %
6. Resting	2036	234	77 %	75 %	76 %
7. Hunt or gather foodstuffs	2036	149	79 %	86 %	82 %
8. Personal care	2036	345	91 %	67 %	77 %
9. Clean the house, wash, or iron	2036	170	79 %	93 %	86 %
10. Fetch water	2036	179	78 %	93 %	85 %
11. Work for a wage or salary	2036	37	89 %	99 %	94 %
12. Running or working in household business	2036	40	85 %	99 %	92 %
Accuracy					77 %
Macro average F_1-score					76 %

The prediction scores demonstrate that our model accurately classifies the reported time use activities in over three-quarters of the cases, using the data from the physical activity sensor signal. To further validate the model's predictive power, we conducted a sensitivity test by varying the dataset size. As the dataset size increases, the prediction scores exhibit a positive trajectory, indicating the model's capacity to improve its performance with growing datasets. This evidence substantiates that time use activities can be reliably predicted using sensor data, effectively addressing our main research question.

For the trained eye **Table 2** also illustrates variations in prediction scores among different activities. Interestingly, despite having balanced train and test sets, the activities less frequently reported prior to the creation of the balanced dataset for analyses (see **Figure 1**) exhibit higher prediction scores. One plausible explanation for these variations is that the more frequently reported activities also originate

from a more diverse group of participants (see **Table 2**), which requires a greater amount of information for building models with accurate predictions.

Another plausible explanation is that the more frequently reported activities often belong to a group of broadly defined time allocation activities with a broader set of sub-activities (or atypical activity motions). For example, the activity 'Social or religious activities and hobbies' can encompass a diverse set of sub-activities and activity motions. Since numerous activities can fall under this broad activity category, it may be the reason why it is the most frequently reported activity, yet the least predictable. Such a diverse group of activity motions may be less accurately captured by purely accelerometer-based sensor devices.

The activities 'Personal care' and 'Fetch water' are cases in point. Despite being reported by a more diverse set of respondents compared to several other activities, such as 'Social or religious', 'Travelling and commuting', or 'Hunt or gather food', they are more confined and well-defined activities, resulting in better prediction scores. In fact, 'Personal care' has a 77 percent F_1-score, surpassing the average prediction, despite having the second highest number of participants.

Both explanations are reasonable, yet hard to disentangle, and it is likely that both may be at play. We can observe explained behavior by visualizing the prediction F_1-scores for each activity against the number of unique participants. In **Figure 2**, the solid line represents the best linear fit (with $R^2 = 0.75$ and p-value = 0.0003) to the data points. The discovered linear relationship underscores the substantial influence of participant diversity on prediction scores. Utilizing this relationship can aid in normalizing prediction scores for underlying participant diversity when comparing activities.

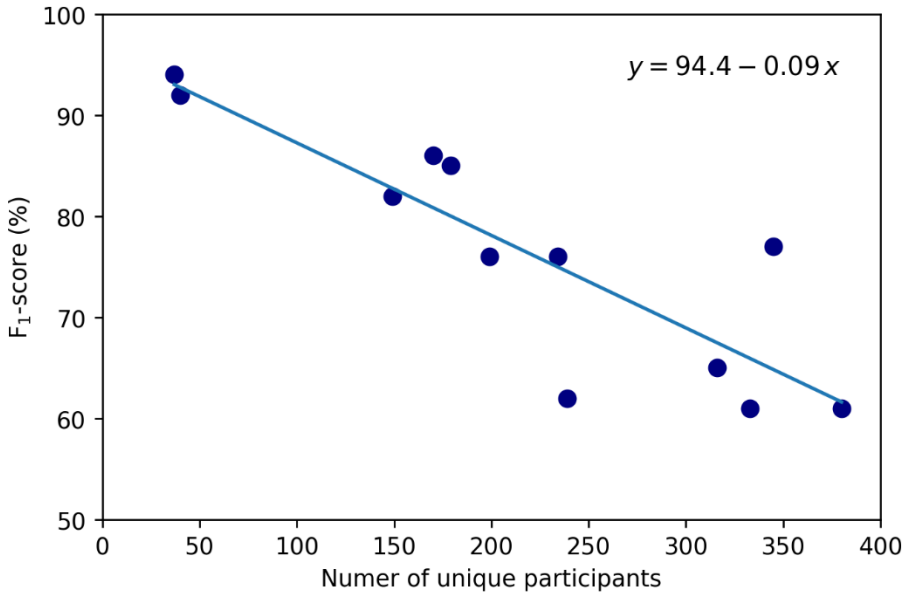

Figure 2. Prediction score versus participant diversity. The best linear fit to the data points is displayed by the solid blue line, where the fitting function is given in the top right corner.

4.2 Common misclassifications

To investigate whether and how the predicted activities are confused with others, we conducted a misclassification study. This aims to identify and comprehend the frequent misclassifications among activities. The misclassifications are illustrated in **Figure 3** through a confusion matrix. The figure should be read from left to right to understand how actual activities are predicted. Darker color shades correspond to more frequent occurrences in the cells.

As expected, the majority of observations lie along the diagonal – indicating accurate predictions. Off-diagonal darker cells indicate combinations of activities that are more likely to be classified differently

in the time use data versus the predicted data. Despite offering insight into the nature of misclassifications, it remains uncertain whether these classification errors stem from the model itself or are a result of the inherent imperfections of the time use data.

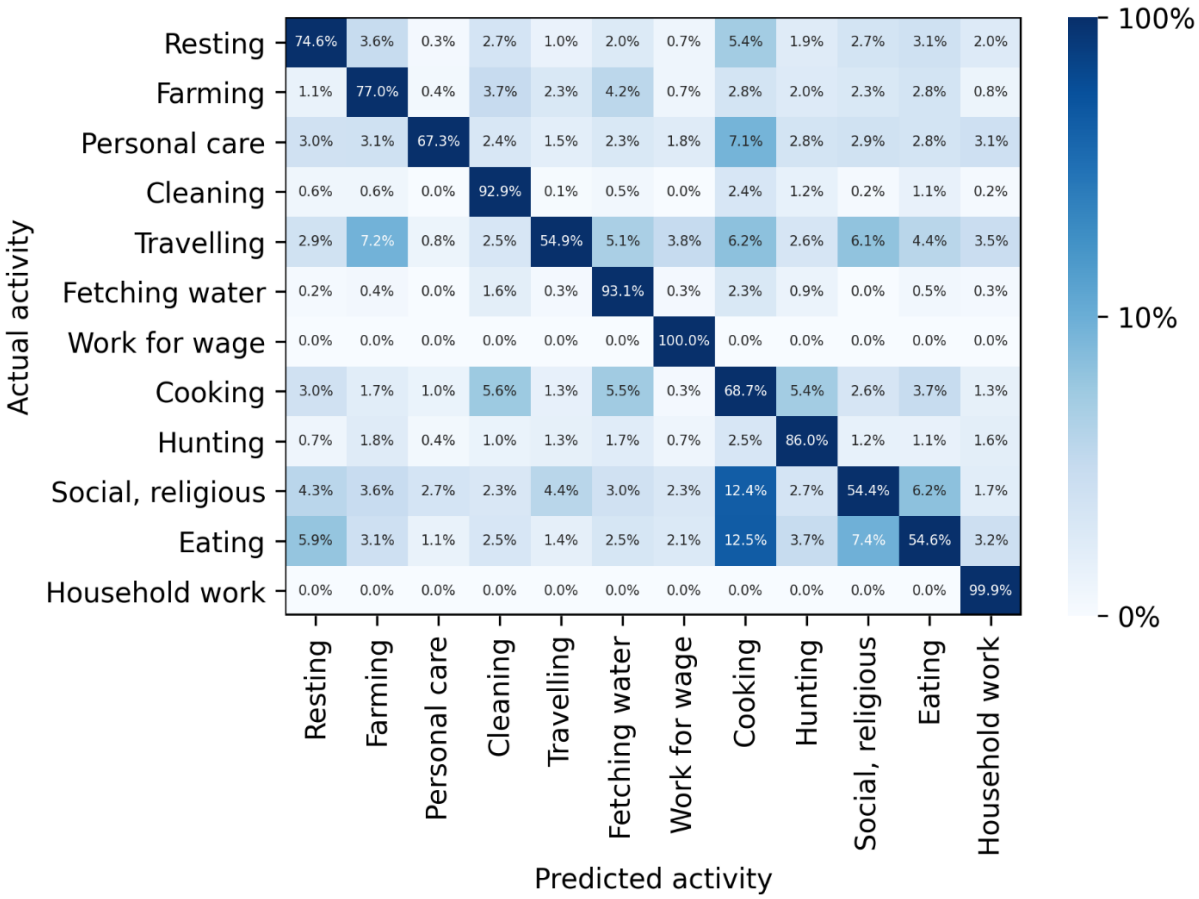

Figure 3. Confusion matrix for the twelve major activities.

The figure illustrates that certain activities, such as 'Work for a wage', and 'Household work' are predicted accurately with minimal confusion, while activities like 'Travelling', 'Eating', and 'Social, religious' are more frequently misclassified. This can be partly attributed to the fact that the latter activities involve more sedentary behavior, making them challenging to differentiate using a waist-worn sensor device. Notably, the activity 'Social, religious' is the most frequently confused activity. Given that it is also the most frequently reported activity in the time use survey, it may serve as a default selection in the questionnaire's time use section when respondents are uncertain.

The activity 'Eating' is mostly misclassified as 'Cooking', while 'Cooking' itself is frequently misclassified as 'Hunting', 'Fetching water', or 'Cleaning'. These misclassifications may be caused by misreporting or by mistiming. Activities performed in succession may be assigned to adjacent time intervals in the recall data. Even a slight shift of a few minutes in the reported activity time-window of 15 minutes during the time use survey can lead to misclassification errors, especially if a shift occurs systematically, such as when activities follow a specific sequence, like fetching water before cooking and cleaning or traveling followed by farming.

4.3 Age and gender-wise activity prediction

We examined the differences of model performance in terms of gender (male versus female) and age (<30 years versus 30+) for the twelve major activities. Based on the prediction scores, some activities are more accurately predicted for specific population groups, although differences are typically small

in most cases. This leads us to question whether certain population groups exhibit characteristic traits in performing activities. For example, it may be that men perform cleaning in a very monotonic manner, making it easier to predict the activity for men. However, other factors also need to be considered before drawing definite conclusions. The prediction scores for the mentioned population groups are presented in **Table 3**.

Table 3
Gender specific model performance per activity.

Activity	Overall		Male		Female		Young (< 30 yr.)		Old (≥ 30 yr.)	
	F_1-score	Np	F_1-score	Np	F_1-score	Np	F_1-score	Np	F_1-score	Np
1. Social or religious activities	61 %	332	67 %	152	50 %	180	62 %	157	60 %	175
2. Farming	76 %	199	79 %	101	72 %	98	74 %	81	77 %	118
3. Cook or prepare food	62 %	239	68 %	34	61 %	205	64 %	113	60 %	126
4. Travelling and commuting	65 %	317	66 %	146	64 %	171	66 %	134	64 %	183
5. Eating and drinking	61 %	379	61 %	167	60 %	212	65 %	173	56 %	206
6. Resting	76 %	234	76 %	102	76 %	132	71 %	113	78 %	121
7. Hunt or gather foodstuffs	82 %	149	87 %	54	80 %	95	85 %	60	81 %	89
8. Personal care	77 %	345	77 %	149	78 %	196	80 %	162	75 %	183
9. Clean the house, wash, or iron	86 %	170	94 %	30	84 %	140	86 %	86	85 %	84
10. Fetch water	85 %	179	95 %	21	84 %	158	87 %	90	83 %	89
11. Work for wage or salary	94 %	37	93 %	28	96 %	9	96 %	14	93 %	23
12. Working household business	92 %	40	92 %	21	92 %	19	92 %	17	91 %	23
All activities	76 %	388	79 %	171	75 %	217	77 %	175	75 %	213

When evaluating population-specific prediction scores, one needs to keep in mind that the set sizes (minutes), and especially the underlying diversity of participants, as indicated by Np (number of participants), can significantly impact the scores. As previously demonstrated in **Section 4.1** with **Figure 2**, there is a noticeable correlation between prediction scores and participant diversity.

To isolate the effects of participant diversity, we apply a size correction to the population-specific prediction scores according to the earlier found linear approximation formula. Since the found relationship with participant diversity (slope) scales as 0.09, for comparing purposes, we adjust the obtained scores to a scaled score as follows:

$$Scaled\ score = F_1\ score + 9 \times \frac{Np}{100}\ (\%)$$

After rescaling, the differences in prediction between the population groups diminish. While it initially appeared that certain activities, such as cleaning for males, were significantly more predictable for a certain gender or specific age groups, the adjusted prediction scores reveal that there is not a substantial difference in prediction accuracy between these population groups. The scaled F_1-scores for gender and age are graphically represented per activity in Figure 4.

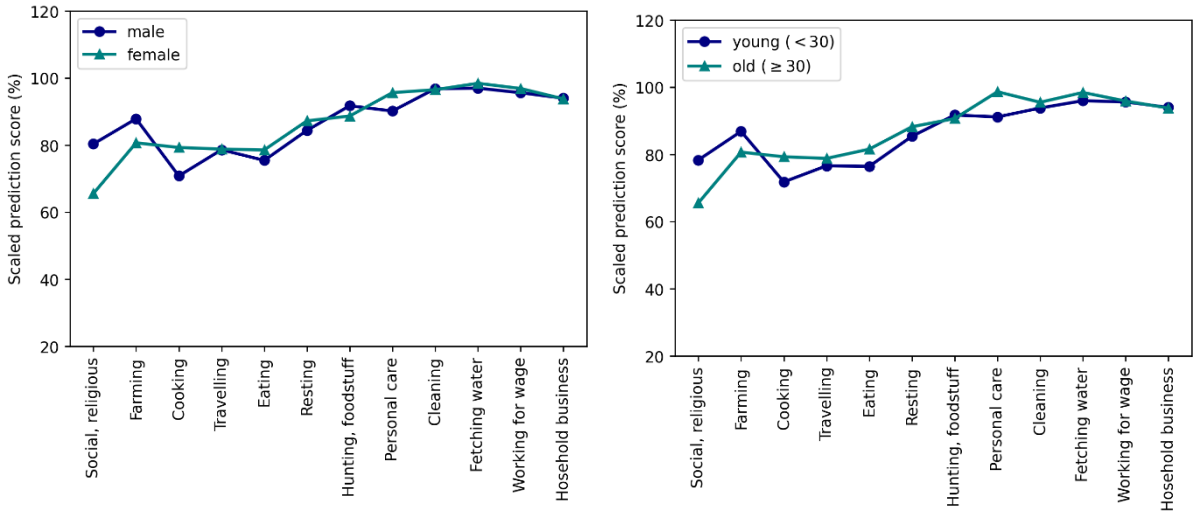

Figure 4. Gender-wise (left) and age-wise (right) scaled prediction scores of activities.

4.4 Impact of background variables on model performance

In this section, we explore the impact of background variables on model performance. We conduct three model runs: one using only sensor data, another using only background variables, and a third considering both datasets. Our findings indicate that incorporating background variables significantly improves model performance. **Table 4** summarizes the model performances for the three conditions.

Table 4
Model performance scores for different sets of variables included in the model.

Data model	F_1-score	Train accuracy	Test accuracy
Full model	76 %	84 %	77 %
Only sensor signal features	64 %	76 %	65 %
Only background features	52 %	56 %	55 %

This outcome suggests that there may be an underlying tendency towards profiling user behavior, especially considering our processed dataset comprises a modest group of 388 participants. In simpler terms, there is a possibility that our model is overfitting the data. To assess the validity of this hypothesis, we compare the model performance by examining their accuracies on both the train and test datasets. When the accuracies diverge significantly, it can be an indication of overfitting. To mitigate this, we applied an early stopping criterion during modeling, using test and train errors.[5] As a result, the train and test accuracies no longer diverge significantly. Unsurprisingly, the models perform less effectively for both the pure sensor signal data and the pure background data, albeit with less overfitting.

[5] The errors are 100 percent minus the accuracy scores. The ratio of test over train error is used in the model through early stopping, with a maximum allowed ratio of 1.5.

To further assess the predictive power of the three models, we vary the dataset size used for model training. If a model has any predictive power, one expects an improving trend in model performance (i.e., F_1-score) with increasing dataset size. The combined plot for assessing predictive power for the three data models is depicted in **Figure 5**.

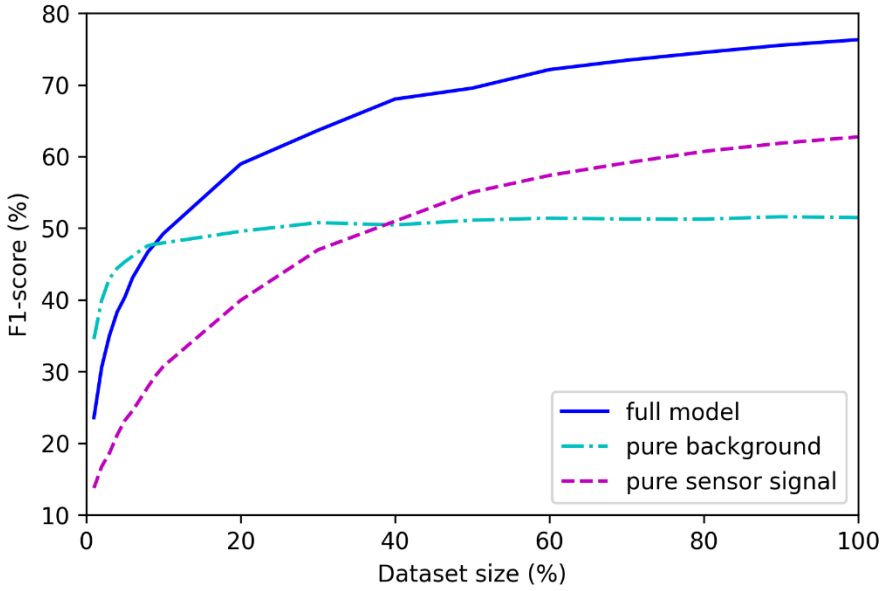

Figure 5. The predictive power of the three data model types.

The figure demonstrates that the model based solely on background variables exhibits minimal improvement with an increasing dataset size, consistently producing an approximate 50 percent base prediction. It does not learn from additional data. We attribute this stagnant performance to the fact that background variables are primarily useful for broadly categorizing activities. Certain activities are more specific to gender or age, or are performed at specific times or days, making it possible to group activities at a higher level. This grouping facilitates learning for a model. If the modeling were tailored to individual profiles, leveraging only background information might potentially yield significantly improved prediction scores.

We observe that when exclusively using sensor signal variables, the optimal score of 64 percent is achieved with the entirety of the available data. While the curve continues to ascend, its slope is diminishing. Nevertheless, the model's performance could still enhance with the addition of more observations.

The full data model experiences a consistent enhancement in performance due to the inclusion of background variables. When observing the left portion of the chart, up to a specific fraction of data, the full data model does not outperform the pure background data model. This means that with limited data, one is better off grouping activities from just background information of participants rather than attempting direct predictions. The shift to improved predictions occurs when the dataset reaches 10 percent of its size, which amounts to 645 hours of logged activity.

Due to diminishing returns, we reach the prediction score of 90 percent of the documented final score (F_1-score of 76 percent) at a dataset size of merely 40 percent of total. This can considerably reduce the need for time use data collection for future studies involving physical activity tracking devices.

5 Model application

Since the physical activity tracking ran for two consecutive weeks, and the time use survey was performed on two separate days, we have 12 days of sensor-based data per person without the accompanying time use survey that we can use to provide insight into people's daily activities. We applied the trained predictive model to the remaining physical activity tracking data to obtain the (modeled) daily time allocation for the complete two-week period.

There are two objectives for doing this. First, by running the model on the complete data, we can validate the model on a global scale by examining the distribution of activities. This will allow us to assess the model's performance qualitatively, which is valuable since the 'true' time use to validate the model against was not observed. Second, by estimating daily activities for a two-week cycle, we can gain insight into activity patterns and gender and age disparities in time allocation towards paid and unpaid activities.

5.1 Model validation

We validate the model by the relative distribution of individuals' time allocation obtained from predictive modeling with the distribution derived from the time use survey. This serves as an additional qualitative validation for assessing the model's effectiveness. The model evaluation performed on the test set (**Section 4.1**) provides a numerical assessment of the model's performance with respect to the time use data. However, it is worth noting that the time use data may contain measurement errors (recall-based) and may not fully represent the 'ground truth'. Therefore, we use this qualitative validation to identify any systematic aberrations in the results. **Figure 6** depicts the percentage-based distribution of people's time allocation as per the time use data and the predictive model.

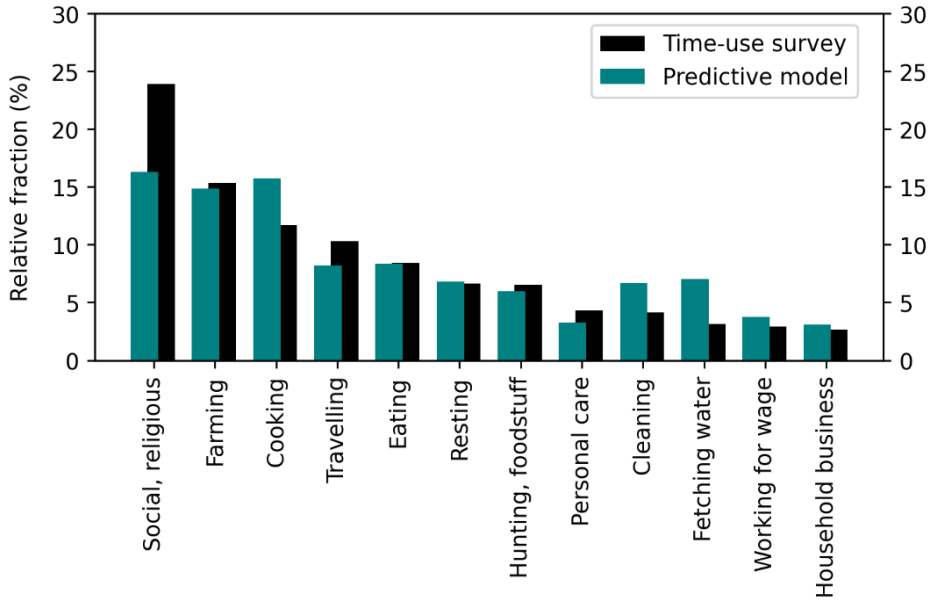

Figure 6. Relative distribution of time spent by activity: time use survey versus model prediction.

There appears to be a strong match in the distribution of activities. A statistical significance test was performed to see if the two datasets are sampled from the same distribution. The two-sided Welch's t-test indicates that the two distributions do not significantly differ from each other (p-value = 1.0), thereby confirming the model's effectiveness. There are, however, noticeable differences in the activities 'Social, religious', 'Cooking', 'Cleaning', and 'Fetching water'. These mismatches align with the common misclassifications highlighted in **Section 4.2**. The predicted fraction of the activity 'Social, religious' is lower, compensated by a higher predicted time share for the 'Cooking', 'Cleaning', and

'Fetching water' activities. Determining the true value is challenging, as participants may also overreport the activity 'Social, religious' in the time use survey.

A direction for future data collection would be to subdivide the activity "Social, religious" into multiple categories, providing a clearer identification of the activities within this currently rather broad activity category. This approach may help reduce ambiguity in reporting social and religious activities, consequently improving predictions for all activities.

5.2 Time allocation by age and gender

The gender-specific relative activity distribution, based on the full model predictions, is presented in **Figure 7**. When examining the daily time allocation of participants, noticeable gender-based differences emerge in the distribution of performed activities. In this rural Malawi sample, the activities 'Social and religious', 'Farming', and 'Working for wage', are predominantly performed by males in this, while activities 'Cooking', 'Cleaning' and 'Fetching water' are primarily performed by females. Model predictions align with the patterns observed in the time use survey data.

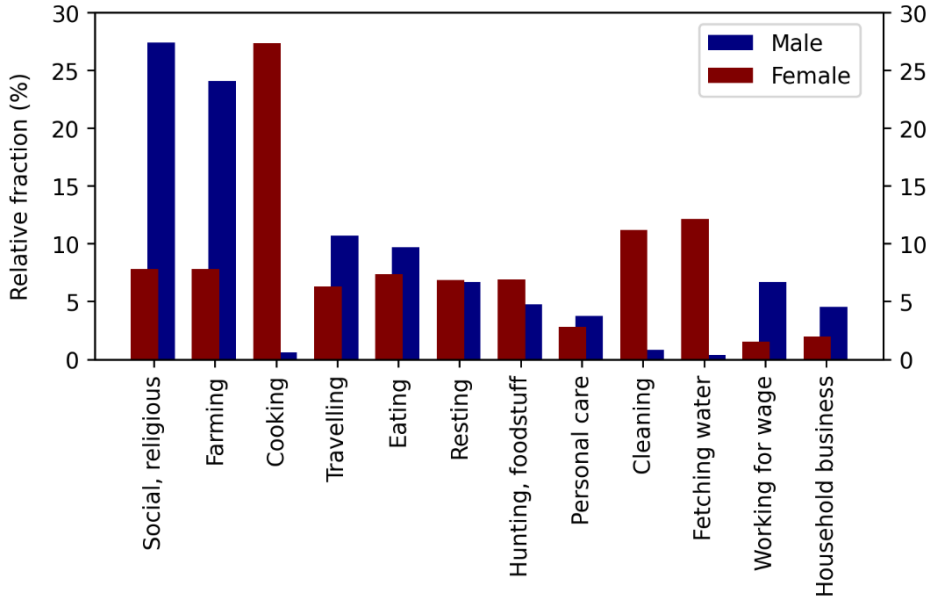

Figure 7. Gender-wise distribution of activities performed.

Similarly, we present the age-wise distribution of the daily time allocation in **Figure 8**. While the age-wise differences are less pronounced as compared to the gender-wise differences, they are still evident. The activities 'Personal care', 'Cleaning', 'Fetching water', and 'Working for wage' are more frequently performed by younger people (< 30 years). These activities demand greater physical strength which younger individuals typically possess, and as a result, they spend more time on them. Conversely, the data indicates that older individuals (≥ 30 years) allocate more time to 'Farming', 'Resting', and 'Hunt or gather foodstuff'.[6]

[6] Additional descriptive statistics (mean and variance) on the time allocation for various population groups in Malawi are provided in Appendix C.

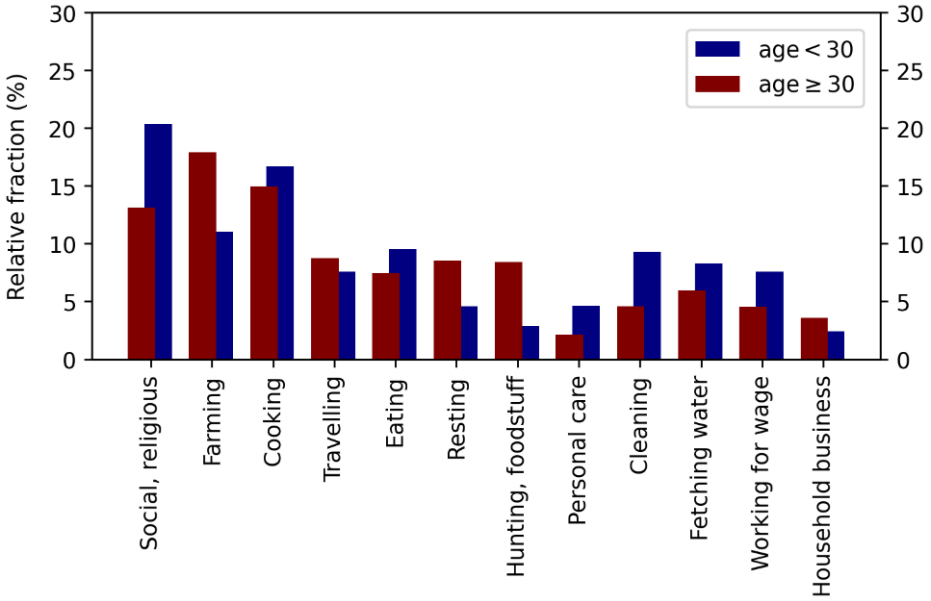

Figure 8. Age-wise distribution of activities performed.

6 Conclusions

The analyses presented in this paper show that we can effectively combine low-cost accelerometer data with recall-based survey data on time use to train a machine learning model that predicts time allocation to 12 major activities with a high degree of accuracy. This holds true despite the limitations imposed by recall-based time use data, the temporal resolution of sensor-based physical activity tracking data collection, and the broad definitions of activities considered in the time use module used in this work. The analyses reveal that the best performing model reaches 90 percent of its ultimate model score with 40 percent of training data. This suggests that the costly and burdensome survey data collection on time use can potentially be omitted in 60 percent of our sample, provided that sensor-based physical activity data collection is extended to the entire sample and our model can provide imputed estimates of time allocation to major activities for the sub-sample not subject to time use data collection.

While the results are highly promising, they may not on their own be sufficient to justify the widespread adoption of accelerometers in large scale surveys, even for a small subsample. This is due to the added logistical complexities and the requisite technical capacity in deploying and managing accelerometers. Simultaneously, when considering the use of these technologies in surveys, one should also consider the additional value they can provide in other domains of information, such as valuable physical activity data that can greatly benefit public health monitoring and food security analysis. These technologies can enable more accurate estimates of individual dietary energy requirements, for example. It is only by weighing all potential uses against the associated costs and burden that survey practitioners can comprehensively establish the costs and benefits of adopting accelerometers in field-based data collection.

Furthermore, the second contribution of our paper is to demonstrate that some activities are harder to predict and classify than others. Potential factors underlying worse predictions are (a) using broad activity definitions that do not sufficiently discriminate across different physical activity levels of sub-activities, and (b) recall bias in time use survey data – resulting in misreporting or mistiming of reported activities. An analysis of the misclassification revealed that the most frequently reported activity ('Social, religious') is also the most often wrongly predicted one, probably because this activity

constitutes too large a bucket including several, diverse activities. We found that some consecutively performed activities, such as fetching water, cooking, eating, and cleaning are often misclassified with each other. Since the motion patterns belonging to these activities should not be similar, the most logical conclusion is that the reported activities are overlapping in time. A key finding here is that misclassifications do not always behave symmetrically. Eating, for example, is often confused with cooking, but the opposite (cooking being confused with eating) is less common. This can be indicative of a systematic (mis)timing offset in one direction.

We evaluated potential differences by age and gender as well. In terms of predictions, we found that there are no significant differences by age or gender, after correcting for any size dependence intrinsic to the data. This suggests that the model is unbiased with respect to gender or age. When applying the model on the whole dataset, we see we pick up noticeable gendered differences in time allocation, which align well with the observed time use data.

One final interesting discovery is that background information, such as participant's age and gender and information on the time of day and the day of week, plays a role in improving model predictions. By running multiple models with different subsets of the data as input, that is, pure sensor signal data, pure background data, and the complete data we were able to disentangle the impact of background information on model predictions. The analysis shows evidence of a fixed boost in model performance when background information is used in combination with the physical activity sensor signal data.

We conclude that it is feasible to predict time use data with high accuracy and that obtainable improvements in the quality of both time use and physical activity data collection hold promise of significantly elevating the predictive capabilities of the underlying models. Based on these results, we offer a few recommendations for the design of future data collection campaigns and suggest some priorities for future research to advance the agenda toward scalable adoption of these techniques for the integrated use of sensor data and time use survey data.

First, to reduce the possibility of misclassification it will be important to revisit the activity lists in time use modules such that the activities with broad definitions ('Social, religious' activities for instance) are split into multiple subcategories. Not only would this avoid ambiguity in reporting social and religious activities, thereby reducing misreporting of activities, but it would also improve the prediction of other activities. Inclusion of more disaggregated activity lists in recall-based time use modules are relatively costless exercises - since the interviewers aim to elicit respondent narratives and classify the activities in accordance with the narrative, rather than asking whether each respondent performed each potential activity in a list for a given time interval. The development of a more disaggregated activity list, however, may involve an initial investment and subsequent refinements, based on qualitative work and additional piloting. For instance, the current International Classification of Activities for Time-Use Statistics (ICATUS) is updated as of 2016 - comprising 165 activity groups, classified into 56 divisions and 9 major divisions – and still represents a manageable number of categories (UN Women, 2021). By comparison, the activity list used in this study included only 26 activity groups.

Second, a key source of misclassification, respondents' recall, can be averted only by stepping away from recall-based time use surveys. Mobile phone apps that allow respondents to log activities in real time, particularly those with activity illustrations that are fit for low-literacy settings, offer promise to reduce recall bias (Daum et al., 2019; Kilic et al., 2024). If implemented well, mobile phone-based time use diaries, coupled with higher-frequency physical activity tracking data collection (i.e., at sub-minute data collection frequency, without additional data collection cost), can potentially improve the predictive power of machine learning models.

Future research in this area should document the predictive accuracy gains, if any, associated with using mobile phone-based time use diaries and sub-minute-level physical activity tracking data collection vis-à-vis relying on traditional, recall-based time use diaries and minute-level sensor data collection as in the case of our current paper. Fortunately, a follow-up study implemented in 2022/23 in Malawi (Kilic et al., 2024) - whose design has been informed by the analysis presented here - will

provide an opportunity to fill this knowledge gap. Leveraging this new data source, future research should also demonstrate whether the minimum-required training data on time use varies by whether the information is elicited through mobile phone- or recall-based time use diaries. Additionally, it will be important to replicate similar studies in different country contexts to understand the extent to which the predictive ability of the model varies in populations with different time use patterns and physical activity profiles.

References

Bell, A., Ward, P., Tamal, M.E.H. et al. Assessing recall bias and measurement error in high-frequency social data collection for human-environment research. Popul Environ 40, 325–345 (2019).

Beneria, L. The enduring debate over unpaid labour. International Labour Review 138(3), 287-309 (1999).

Breiman, L. Bagging predictors. Machine Learning, 24, 123-140 (1996).

Breiman, L. Random Forests. Machine Learning, 45(1), 5-32 (2001).

Chandler, J. L., Brazendale, K., Beets, M. W. & Mealing, B. A. Classification of physical activity intensities using a wrist-worn accelerometer in 8-12-year-old children. Pediatr. Obes. n/a–n/a (2015).

Dieu, O. et al., Physical activity using wrist-worn accelerometers: comparison of dominant and non-dominant wrist. Clin. Physiol. Funct. Imaging, n/a–n/a, (2016).

Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al., Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLoS ONE 12(2): e0169649. (2017).

Ellis, K., Kerr, J., Godbole, S., Staudenmayer, J. & Lanckriet, G. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification. Med. Sci. Sports Exerc. 48, 933–40 (2016).

Fafchamps, M., Kebede, B. & Quisumbing, A. R., Intrahousehold welfare in rural Ethiopia. Oxford Bulletin of Economics and Statistics 71(4), 567–599 (2009).

Frankin, B. (1748). Advice to a Young Tradesman. *Founders Online,* National Archives. [Original source: *The Papers of Benjamin Franklin,* vol. 3, *January 1, 1745, through June 30, 1750*, ed. Leonard W. Labaree. New Haven: Yale University Press, 1961, pp. 304–308.]

Gaddis, I., Oseni, G., Palacios-Lopez, A. & Pieters, J. Who is employed? evidence from sub-Saharan Africa on redefining employment. Journal of African Economies 32(2), 151-174 (2023).

Folbre, N. The *International Labour Review* and gender equality: the importance of women's unpaid and paid work. International Labour Review 160(4), e43-e51 (2021).

Friedman, J., Gaddis, I., Kilic, T., Martuscelli, A., Palacios-Lopez, A. & Zezza, A. The distribution of effort: physical activity, gender roles, and bargaining power in an agrarian setting. The World Bank Economic Review 37(1), 93-111 (2023).

Hildebrand, M., VAN Hees, V. T., Hansen, B. H. & Ekelund, U. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. *Medicine & Science. Sports* & Exercise. Volume 46, Pages 1816–1824 (2014).

International Labour Organization. Resolution I. Resolution concerning statistics of work employment and labour underutilization. Retrieved from http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_230304.pdf (2013).

Kamada, M., Shiroma, E. J., Harris, T. B. & Lee, I. M. Comparison of physical activity assessed using hip- and wrist-worn accelerometers. *Gait & Posture, Volume* 44, Pages 23–28 (2016).

Kapteyn, A., Banks, J., Hamer, M., Smith, J.P., Steptoe, A., Van Soest, A.H.O., Koster, A., Saw H.W., "What They Say and What They Do: Comparing Physical Activity Across U.S., England, and the Netherlands", Journal of Epidemiology & Community Health Volume 72, Pages 471–476(2018).

Kilic, T. Koolwal, G., Drazi Vundru, W. et al. Recording the time divide: a comparative study of smartphone and recall-based approaches to time use measurement. World Bank Policy Research Working Paper No. 10695 (2024).

Killingsworth, M. R. & Heckman, J. J., Chapter 2: Female labor supply: a survey. Handbook of Labor Economics 1, 103-204 (1986).

Mannini, A., Intille, S. S., Rosenberger, M., Sabatini, A. M. & Haskell, W. Activity recognition using a single accelerometer placed at the wrist or ankle. Med. Sci. Sports Exerc. 45, 2193–203 (2013).

Menai, M., van Hees, V., Elbaz, A. et al., Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. Scientific Report 7, Pages 45772 (2017).

Migueles, J. H. et al., Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. Sport. Med. 47, 1821–45 (2017).

Nunavath, V., Johansen, S., Johannessen, T. S. et al. Deep learning for classifying physical activities from accelerometer data. Sensors 21(16), (2021).

Pencavel, J., Chapter 1: Labor supply of men: a survey. Handbook of Labor Economics 1, 3-102, (1986).

Pfeiffer, K. A., Lisee, C., Westgate, B. S. et al. Using accelerometers to detect activity type in a sport setting: challenges with using multiple types of conventional machine learning approaches. Measurement in Physical Education and Exercise Science 27(1), (2023).

Pratt, M., Sallis, J. F., Cain, K. L. et al. Physical activity and sedentary time in a rural adult population in Malawi compared with an age-matched US urban population. BMJ Open Sport & Exercise Medicine 6(1), (2020).

Romanzini, M., Petroski, E. L., Ohara, D., Dourado, A. C. & Reichert, F. F. Calibration of ActiGraph GT3X, Actical and RT3 accelerometers in adolescents. Eur. J. Sport Sci. 1–9 (2012).

Shiroma, E. J. et al., Daily physical activity patterns from hip- and wrist-worn accelerometers. Physiol. Meas. 37, 1852–1861 (2016).

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D. & Freedson, P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J. Appl. Physiol. 107, 1300–1307 (2009).

Sundararajan, K., Georgievska, S., te Lindert, B.H.W. et al., Sleep classification from wrist-worn accelerometer data using random forests. Scientific Report 11, Pages 24 (2021).

Wales, T. J. & Woodland, A. D., Estimation of the allocation of time for work, leisure, and housework. Econometrica 45(1), 115-132 (1977).

Wennman, H., Pietilä, A., Rissanen, H. et al., Gender, age, and socioeconomic variation in 24-hour physical activity by wrist-worn accelerometers: the FinHealth 2017 Survey. Scientific Report, Volume 9, Pages 6534 (2019).

Willetts, M., Hollowell, S., Aslett, L. et al., Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Scientific Report 8, pages 7961 (2018).

Zhang, S., Rowlands, A. V., Murray, P. & Hurst, T. L. Physical activity classification using the GENEA wrist-worn accelerometer. Med Sci Sport. Exerc 44, 742–748 (2012).

MODULE : TIME ALLOCATION

PLEASE RECORD A LOG OF THE ACTIVITIES FOR THE INDIVIDUAL IN THE LAST COMPLETE 24 HOURS (STARTING YESTERDAY MORNING AT 4 AM, FINISHING 3:59 AM OF THE CURRENT DAY). THE TIME INTERVALS ARE MARKED IN 15 MIN INTERVALS. MARK ONE PRIMARY ACTIVITY FOR EACH TIME PERIOD BY ENTERING THE CORRESPONDING ACTIVITY CODE IN THE BOX. A SECONDARY ACTIVITY (OPTIONAL) CAN BE ENTERED IN CASE OF SIMULTANEOUS ACTIVITIES.

THIS FORM MUST BE ADMINISTERED TO THE **RESPONDENT HIMSELF/HERSELF** DURING THE **2ND AND 3RD VISITS** TO THE HOUSEHOLD.

Now I'd like to ask you about how you spent your time during the past 24 hours. We'll begin from yesterday morning, and continue through to this morning. This will be a detailed accounting. I'm interested in everything you did (i.e. resting, eating, personal care, work inside and outside the home, caring for children, cooking, shopping, socializing, etc.), even if it didn't take you much time.

	Night		Morning	Day								
	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00
1. Primary Activity (WRITE ACTIVITY CODE)												
2. Secondary Activity (WRITE ACTIVITY CODE)												
	Day		**Evening**	**Night**								
	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00	1:00	2:00	3:00
3. Primary Activity (WRITE ACTIVITY CODE)												
4. Secondary Activity (WRITE ACTIVITY CODE)												

ACTIVITY CODES

A.....Sleeping and resting
B.....Eating and drinking
C.....Personal care
D.....School (incl. homework)
E.....Work for a wage, salary, commission or in-kind payment (incl. ganyu, paid apprenticeships)
F.....Run, work or help in a non-agricultural and non-fishing household business
G...Work for other households free of charge as exchange laborer
H.....Farming
I.....Livestock
J.....Fishing
K.....Hunt or gather foodstuffs
L.....Buy food or other items or obtain services
M.....Make goods (furniture, pottery, baskets, clothing)
N.....Cook or prepare food or drinks to preserve them
O.....Collect firewood or other natural products
P.....Fetch water from natural or public sources
Q.....Clean the house, wash or iron
R..Household maintenance or own construction work (e.g. to renovate, extend or build the household's dwelling)
S.....Provide care or assistance to adults (18+ years)
T.....Look after children (17 years or younger)
U.....Plan the household's finances or bills
V.....Travelling and commuting
W....Watching TV/listening to radio/reading
X.....Exercising
Y...Social or religious activities and hobbies
Z ... .Other

Appendix B: Variables used in the models and their descriptions

Table B1
Detailed list of features in the dataset.

Feature	Origin	Description / derivation formula
X acceleration	Base	The mean acceleration in X direction in one minute (in units of g)
Y acceleration	Base	The mean acceleration in Y direction in one minute (in units of g)
Z acceleration	Base	The mean acceleration in Z direction in one minute (in units of g)
Step count	Base	The number of steps in one minute
Age	Background	Age of the participant
Gender	Background	Gender of the participant
Weight	Background	Weight of the participant
Timeofday	Derived	Is it a weekday or a weekend day
Dayofweek	Derived	Part of the day
Vector magnitude	Derived	$\sqrt{}(x^2 + y^2 + z^2)$, i.e., the Euclidean norm
Angle 1	Derived	atan(x/yz); this is also an approximation of Pitch
Angle 2	Derived	atan(y/xz)
Angle 3	Derived	atan(z/xy)
Angle 4	Derived	atan(z/xz)
Angle 5	Derived	atan(x/y)
Angle 6	Derived	atan(y/z); this is also an approximation of Roll
Angle 7	Derived	atan(z/xyz)
Feature cross 1	Derived	X acceleration^2
Feature cross 2	Derived	X acceleration × Y acceleration
Feature cross 3	Derived	X acceleration × Z acceleration
Feature cross 4	Derived	X acceleration × Step count
Feature cross 5	Derived	Y acceleration^2
Feature cross 6	Derived	Y acceleration × Z acceleration
Feature cross 7	Derived	Y acceleration × Step count
Feature cross 8	Derived	Z acceleration^2
Feature cross 9	Derived	Z acceleration × Step count
Feature cross 10	Derived	Step count^2
Epoch mean 1	Derived	15-minute mean of X acceleration
Epoch mean 2	Derived	15-minute mean of Y acceleration
Epoch mean 3	Derived	15-minute mean of Z acceleration
Epoch mean 4	Derived	15-minute mean of Step count
Epoch mean 5	Derived	15-minute mean of Vector magnitude
Epoch standard deviation 1	Derived	15-minute standard deviation of X acceleration
Epoch standard deviation 2	Derived	15-minute standard deviation of Y acceleration
Epoch standard deviation 3	Derived	15-minute standard deviation of Z acceleration
Epoch standard deviation 4	Derived	15-minute standard deviation of Step count
Epoch standard deviation 5	Derived	15-minute standard deviation of Vector magnitude

Appendix C: Population-wise relative time allocation

Table C1

Percentage of time spent on each activity for selected population groups giving mean (μ) and standard deviation (σ).

Population groups	Gender (%)				Age (%)				BMI (%)						Day of week (%)				Gender and Age type (%)							
	Male		Female		≥ 30 yr.		< 30 yr.		≥ 25		< 18.5		18 - 25		Week		Weekend		< 30 yr. male		≥ 30 yr. male		< 30 yr. female		≥ 30 yr. female	
Activities	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
Clean the house	0.7	2.2	15.2	16.6	6.3	12.1	12.0	16.4	16.0	17.5	7.6	12.0	8.7	14.5	9.0	15.3	9.0	13.6	1.1	2.8	0.3	1.3	20.6	17.6	10.9	14.4
Cook or prepare food	0.5	2.8	26.1	15.5	13.8	17.1	16.1	17.5	19.1	15.6	6.5	9.7	15.6	17.8	14.2	16.8	14.1	20.4	0.9	3.9	0.2	1.1	28.2	14.4	24.4	16.3
Eating and drinking	8.6	7.3	6.2	5.8	5.9	6.2	9.0	6.7	6.4	8.8	9.7	7.8	7.1	6.3	7.0	7.1	7.1	7.3	11.5	7.8	6.2	6.0	7.0	4.8	5.6	6.4
Farming	26.3	21.4	9.5	12.0	20.9	20.2	11.9	15.5	7.5	13.6	21.9	22.5	16.8	18.4	19.1	20.9	19.1	15.6	21.3	17.8	30.5	23.3	4.5	7.5	13.6	13.4
Fetch water	0.4	2.1	12.4	12.7	5.5	10.2	9.0	12.3	8.1	12.2	4.9	8.6	7.3	11.5	7.8	12.3	7.8	9.6	0.6	2.7	0.2	1.5	15.7	12.9	9.6	11.9
Hunt or gather foodstuffs	3.8	9.2	5.3	10.6	7.0	12.6	1.7	3.9	7.0	10.7	4.8	8.5	4.5	10.2	4.0	8.8	4.0	15.3	2.1	4.4	5.1	11.6	1.4	3.4	8.4	13.2
Personal care	2.2	2.9	1.2	2.9	1.0	1.7	2.4	3.8	1.4	1.5	1.6	2.8	1.6	3.0	1.9	3.3	1.9	2.8	3.4	3.3	1.2	2.0	1.6	4.0	0.8	1.4
Household business	4.7	14.8	2.4	9.3	3.9	12.7	2.8	11.2	6.7	14.5	0.2	1.3	3.6	12.5	4.2	14.0	4.1	9.6	2.3	10.4	6.7	17.4	3.2	11.8	1.7	6.4
Resting	6.5	11.7	6.4	12.4	8.1	14.1	4.4	8.6	5.8	14.4	14.6	18.0	5.6	10.9	7.1	13.4	7.1	9.6	4.5	7.3	8.2	14.2	4.3	9.5	8.1	14.1
Social or religious	26.3	18.1	6.5	9.4	12.2	13.1	18.8	20.2	12.7	13.2	23.6	26.3	14.4	15.7	9.6	15.9	9.6	25.1	34.5	19.3	19.5	13.7	6.4	9.4	6.7	9.5
Travelling and commuting	12.9	12.4	6.9	9.2	10.1	11.4	8.9	10.8	8.6	9.0	4.2	6.0	10.2	11.5	10.6	12.1	10.6	10.7	12.7	11.4	13.1	13.3	5.9	9.3	7.8	9.0
Work for a wage	7.1	15.7	1.9	10.2	5.2	14.6	3.0	10.9	0.6	2.5	0.3	1.3	4.8	13.9	5.5	17.1	5.4	5.1	5.2	13.2	8.8	17.4	1.2	8.3	2.4	11.4