A data-driven study on preferred situations for running

Author(s)
Wang, Shihan; Timmers, Joris Alexander; Scheider, Simon; Sporrel, Karlijn; Akata, Zeynep; Kröse, Ben

DOI
10.1145/3267305.3267552

Publication date
2018

Document Version
Final published version

Published in
UbiComp '18

Poster: A Data-driven Study on Preferred Situations for Running

Shihan Wang
Informatics Institute
University of Amsterdam
Amsterdam, Netherlands
s.w.wang@uva.nl

Karlijn Sporrel
Department of Human Geography and Planning
University of Utrecht
Utrecht, Netherlands
k.sporrel@uu.nl

Joris Alexander Timmer
Informatics Institute
University of Amsterdam
Amsterdam, Netherlands
joris.timmer@student.uva.nl

Zeynep Akata
Informatics Institute
University of Amsterdam
Amsterdam, Netherlands
z.akata@uva.nl

Simon Scheider
Department of Human Geography and Planning
University of Utrecht
Utrecht, Netherlands
s.scheider@uu.nl

Ben Kröse
University of Amsterdam and Amsterdam University of Applied Sciences
Amsterdam, Netherlands
b.j.a.krose@uva.nl

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3267552

Abstract
We analyzed a large data set from a mobile exercise application to find the preferred running situations of a large number of users. We categorized the users according to their running behaviors (i.e. regularly active, or rarely active over the year), then studied the influence of 15 features, including temporal, geographical and weather-based features for different user groups. We found that geographical features influence the behavior of less active runners.

Author Keywords
Physical activity; Mobile data analysis; Clustering.

ACM Classification Keywords
H.5.m [Information systems]: Mobile information processing systems

Introduction
Physical inactivity has been identified as a leading risk factor in health. Exercise applications (apps) on mobile phones are seen as a potential leverage for stimulating physical activity, as they can monitor users' behavior throughout the day and deliver motivational interventions in the right situations [3]. However, the right situations (for example time, location, weather) for people to perform physical activity might not always be the same for all individuals [1]. Recent research shows that the right situations interact with individ-

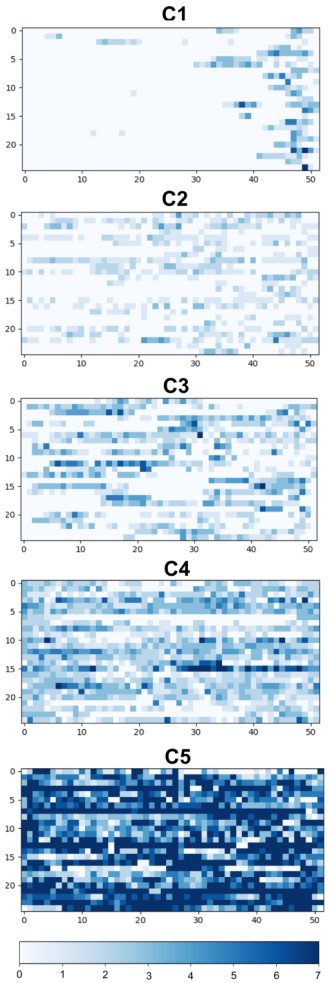

Figure 1: Temporal patterns of 25 random users in each cluster, where the y-axis represents the user and the x-axis represent week of a year. The gray scale implies the running frequency of the user in a week.

ual differences [4]. If such situations could be reliably estimated from data, this would open up enormous potential for adequately timed interventions to trigger physical activities.

In this paper we focus on the running activity. We present a data-driven study using a large-scale dataset, containing about 4 years of running history of over 10K users in the Netherlands, tracked by a smart-phone application. The contribution of this paper is that we are able to characterize runners based on their temporal activity pattern, and study the preferred situations for each group. We analyze temporal, geographical context and weather-related features in each user group, as well as a comparison across groups.

Data Description

We collect a large-scale running dataset by tracking the running activities of over 10K Dutch participants, aged between 18 and 65, while using the MYLAPS exercise app from 2013-03-23 to 2017-03-15. In total, our dataset contains around 440K runs from various users, identified by a unique running ID and grouped by a unique user ID.

Our dataset contains the date, start and end time of the running activity. Furthermore, it contains a set of meta-data, such as the weather, temperature, wind and humidity for each run. Finally, a GPS tracker embedded in the mobile device provides GPS signals, which can be used to extract various geographical context features that might influence the activity. Note that we filter out running activity samples with missing or erroneous values.

Clustering the Users

In this section, we describe the clustering method to distinguish between runners based on their temporal activity patterns. We concentrate on *temporal activeness* which captures user's behavioral characteristics based on the intensity and the regularity of their activity pattern.

Capturing temporal activeness

We format the historical running data of all users into a matrix of activity frequency in the following manner. We start by filtering users with less than 10 running activities in 4 years, since they form the extreme outliers in our dataset. We then characterize the running pattern by measuring intensity as the activity frequency per week, and regularity as an annual activity sequence. The activity sequence captures a consecutive 52 weeks within 4 years in which the user is most active via a sliding window mechanism. In this way, we build a data matrix with 5346 distinct users (i.e.$D \in \mathbb{R}^{5346\times52}$) performing about 280K running activities which accounts for 68% of all their running records.

Clustering users based on temporal activeness

For grouping users, we employ the hierarchical clustering algorithm as follows. First, we measure the similarity between data points via dynamic time warping (DTW). Instead of only comparing an individual value at a certain time index, DTW compares two paired data series by transforming their indices over the entire time period. Furthermore, the Ward variance minimization algorithm is applied to merge two newly formed clusters that are close, by minimizing the variance within the new cluster. In our data, excluding 3 outliers, we find 5 distinctive clusters of users that are referenced as C1 to C5 in this paper.

Visualizing temporal activeness in user groups

We visualize the characteristics of the clusters by randomly selecting 25 users from each cluster in Fig. 1. As summarized in Table 1, our visual analysis clearly discriminates between runners with different patterns, considering both their weakly running intensity and yearly regularity. In the following statistical analysis, we merge the two smallest

List of Features
Temporal:
- Time in a day
- Day in a week
- Week in a year
Weather-related:
- Temperature
- Weather type
- Wind type
- Humidity type
Geographical context:
Distance to ...
- Parks
- Agriculture areas
- Sports areas
- Recreation areas (camping, animal/theme park, playground...)
- Forests
- Nature areas
- Water areas (inland water and coast)
- Traffic areas (street and traffic infrastructure)

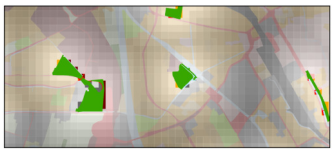

Figure 2: An example of some park areas in Utrecht with computed rasters.

Cluster	Description of activity pattern	Intensity	Regularity	Average run
C1	1167 users with one or few short burst of activity in a year	medium	low	19.2
C2	1503 users with minor activity loosely throughout the year	low	low to medium	28.5
C3	1173 users with minor activity spreading throughout the year	low to medium	medium	61.5
C4	987 users with major activity (3-4 times/week) consistently over a year	medium	high	92.1
C5	26 users with major activity (4-7 times/week) consistently over a year	high	high	160.9

Table 1: A description for temporal activity patterns of captured clusters, including average runs per user, levels of intensity and regularity.

clusters, i.e. C4 and C5 (with 26 users), because they have similar running patterns (users in both clusters run consistently over the entire year).

Features Measuring Preferred Situations

Many features may influence the running patterns of participants [2]. In this study, we combine three kinds listed on the left, i.e. temporal ones that define the calendar time, geographical context ones that define geographic landscape and weather-related ones that define weather conditions. For our temporal features, the calendar time information was derived from meta-data of 'timestamp at start-point'.

Given the GPS point of the start location of the activity, we capture the geographical context of a running event by computing its spatial distance to the nearest region of various landuse classes using an external dataset of landuse from the Dutch Central Bureau for Statistics (CBS) [1]. It contains a collection of spatial regions with various homogeneous landuse classes. We reclassify the given CBS classes (as indicated on the left).

In a second step, we generate a regular 100 m^2 grid over the entire Netherlands to enrich GPS data with corresponding landuse distance information [2]. We compute the linear spatial distance (in meter) from where the user starts running to the nearest region for each of the selected landuse categories (see the example for 'parks' in Fig. 2), and use them as our geographical context features.

Analyzing Preferred Running Situations

In this section, we address the preference towards individual features in each user group, as well as across different user groups. We visualize the preferred patterns in each user group via a set of histograms of feature distribution. Furthermore, via Kolmogorov−Smirnov (KS) we test whether a given pair of feature distributions between two user groups can be considered equal (H_0) or not (H_1).

We notice that runners from different groups share preferences for weather-related situations, where average p-values are around 0.5 and usually above a threshold of 0.05 (thus H_0 should be accepted). They prefer to run at a temperature around 5 to 18 degree, with sunny or half cloudy weather, light wind and high humidity.

Similar patterns across groups appear for temporal features (see Fig. 3): 1) there are two running peaks in the year in

[1] *Bestand Bodemgebruik* (BBG 2012) from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische%20data/natuur%20en%20milieu/bestand-bodemgebruik

[2] Amersfoort / RD New projection: http://spatialreference.org/ref/epsg/amersfoort-rd-new/

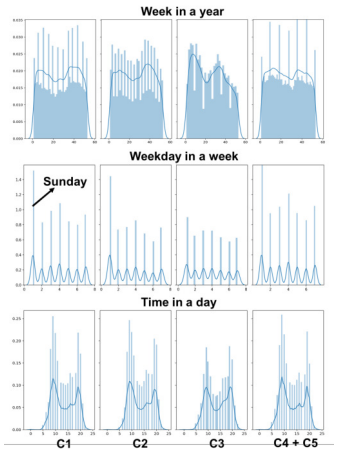

Figure 3: The normalized distribution of temporal features in 4 user groups.

Feature	C1	C2	C3	C4/5
park	4098	4834	5017	5128
agric	3676	4505	4532	4769
sport	3848	4634	4764	4890
recr	3720	4470	4522	4670
forest	3757	4504	4568	4716
nature	3525	4292	4313	4511
water	2808	3497	3715	4004
traffic	3510	4296	4355	4520

Table 2: The standard deviation of distance in meter from where the run starts to the nearest region of a certain landuse category in 4 user groups (agriculture and recreation is shorted as 'agric' and 'recr').

March (winter over time), September and October (time after the summer holiday), to avoid too cold or hot temperatures. 2) likely due to their working schedule, runners generally prefer Sundays and either early mornings or nights.

Although a majority of runners start within few landuse categories (for instance, about 50% running activities are performed within agriculture places, which are the dominant landuse type in the Netherlands), we observe a significant diversity for all spatial context features illustrating different preferences of user groups. The largest p-value of KS tests between all user group pairs is below 0.001 (accepts H_1) in each case.

To further analyze, we compute the mean and standard deviation of distances for each feature in each individual user group. For all kinds of geographical context features, an increase in standard deviation is apparent in Table 2, from the user group with the least active pattern (C1) to that with the most active pattern (C4+C5). This result indicates that runners with more active running patterns have a less clear preference for the chosen landscape situations. At the same time the mean also increases in the same way, indicating that more active users on average run further away from the chosen landscape areas. Those two observations together imply that for more frequent runners the vicinity of the chosen landscape situations is less important, while the opposite is true for less frequent runners. This provides empirical evidence that landscape features could be important to motivate less active citizens towards a healthier lifestyle, as they belong to a user group that is more easily affected by environmental elements.

Conclusion and Discussion

Our findings indicate that although different types of runners share preferences for weather-related and temporal situations, they show different preferences for geographical areas. In general, less active runners prefer stimulating landscape situations. Those findings open up the potential for effectively persuading people to engage in physical activity by timed interventions. In our future work, we plan to develop a prediction model and which can be implemented in an app that provides interventions in optimal situations to motivate less active people towards a healthier lifestyle.

Acknowledgements

We especially thank our cooperator MYLAPS for the dataset. This work is funded by Playful Data-driven Active Urban Living project under NWO and SIA grant 629.004.013.

REFERENCES

1. Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American psychologist* 37, 2 (1982).
2. Nancy Humpel, Neville Owen, and Eva Leslie. 2002. Environmental factors associated with adults' participation in physical activity: a review. *American journal of preventive medicine* 22, 3 (2002).
3. Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2017).
4. Gaurav Paruthi, Shriti Raj, Natalie Colabianchi, Predrag Klasnja, and Mark W Newman. 2018. Finding the Sweet Spot (s): Understanding Context to Support Physical Activity Plans. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018).