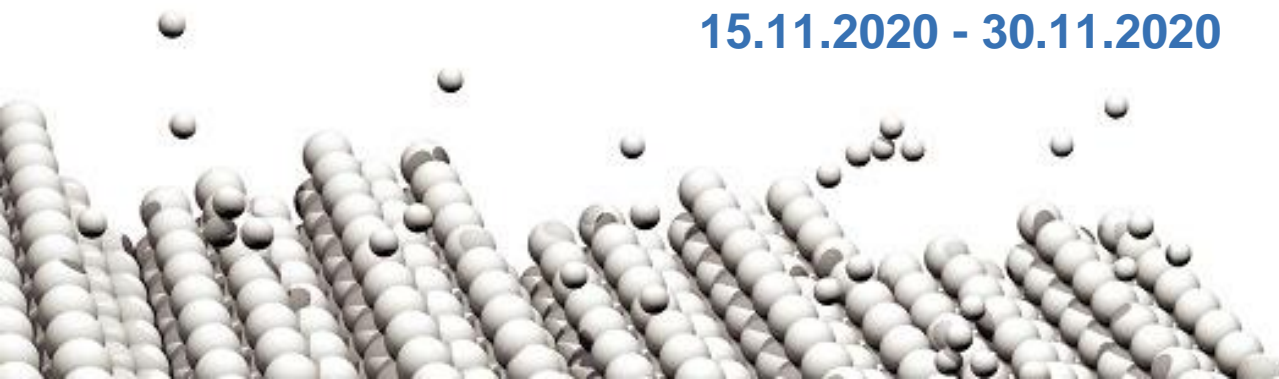# Low-cost environmental and motion sensor data for complex activity recognition: proof of concept

Rok Novak, David Kocman, Johanna Amalia Robinson, Tjaša Kanduč, Denis Sarigiannis, Sašo Džeroski and Milena Horvat

EXPOSURE

PM CONCENTRATIONS → **PM INTAKE DOSE**

- Exposure to airborne particulate matter (PM) can be dependent on a specific activity

- Aggregating this data according to specific activities can provide individuals with detailed information

- They can more effectively reduce their exposure and intake dose by curtailing their activities

How to obtain data about a specific activity at each moment?

Pen&paper? An app?

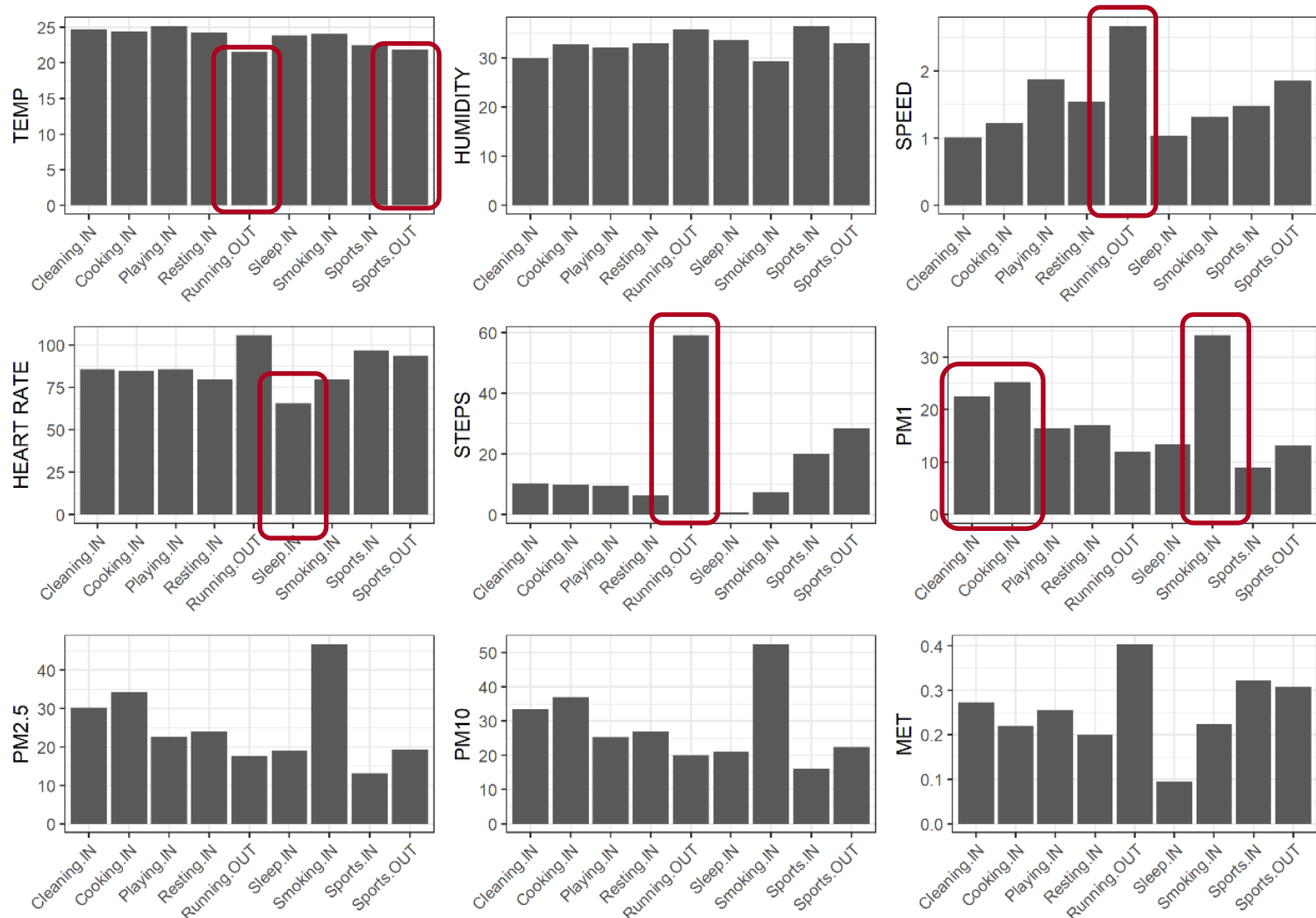*Inaccurate and time consuming*

Why not try machine learning?

Jožef Stefan Institute

Smart activity tracker

Heart rate

Movement

Portable PM sensor

PM

Temperature

Humidity

Speed

Time Activity Diary

Activities + duration

Transport

Conditions in room

- 97 participants were equipped with the two sensors and the TADs (Time Activity Diaries) for 7 days and two seasons
- Manually recording their activities for each hour of the day

Cleaning the data

Analyzing the data

Using the data in three classification models

Jožef Stefan Institute

Data analysis

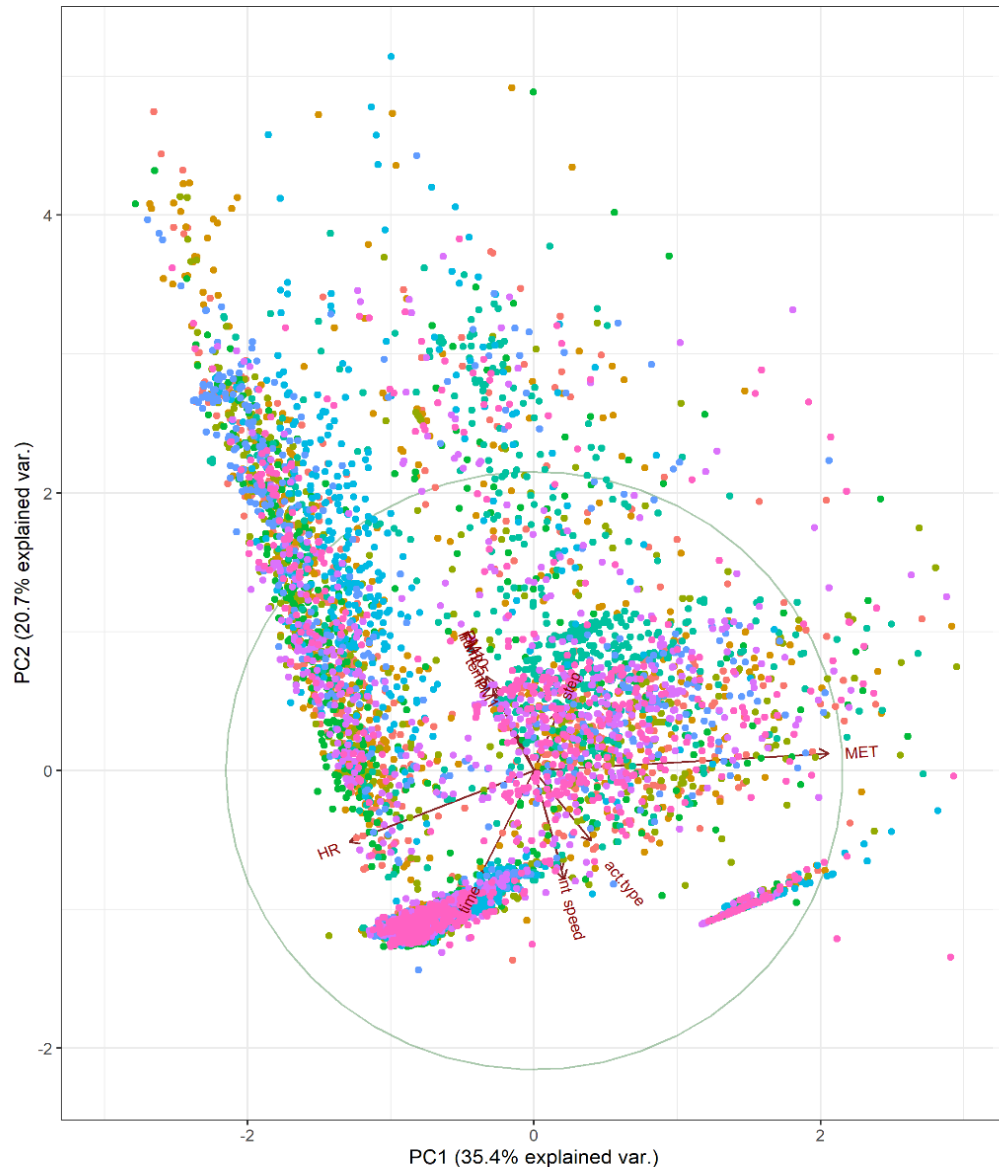Figure 1: Average values for all variables and activities

Figure 2: PCA all variables, activities grouped by color

- PCA shows very few patterns for specific activities:
  - Clustering of running in upper right corner, positive association with steps and movement
  - Some clusters mostly indicating frequent zero values (steps, speed)

- The average values do provide some insight:
  - High PM values for smoking, cooking, cleaning
  - Fast movement and high heart rate for running
  - Lower temperatures for outdoor activities
  - Lower heart rate and movement for sleeping

Classification

- Three classification algorithms were chosen, based on best practices and recommendations.

Classifier	Description
IBk [26]	Instance Based learner otherwise known as k-nearest neighbor (kNN) classifier; selects value of k based on internal cross-validation.
J48 [27]	J48 is a Java implementation of the C4.5 decision tree algorithm developed in 1993 by Ross Quinlan [18]. It can be used for classification and allows a high number of attributes. Deemed as "machine learning workhorse", ranked nr. 1 in the Top 10 Algorithms in Data Mining [28].
RandomForest [29]	Constructs a forest of decision trees in a randomized manner. Developed by Leo Breiman in 2001 [30].

Jožef
Stefan
Institute

Results & discussion #1

Classifier	Correctly classified	Kappa	True Positive	False Positive	Precision	ROC area	PRC area
IBk	32.7%	0.2424	0.327	0.084	0.363	0.621	0.220
J48	39.5%	0.3195	0.395	0.076	0.407	0.767	0.370
Random Forest	43.1%	0.3601	0.431	0.071	0.432	0.807	0.444

- All the used classifiers showed accuracy above 30%, with RandomForest being the most accurate with 43.1%

- As the labeled data was made from hourly labeled activities, this gives it less resolution and more errors (some activities don't last an hour, and most don't last exactly a set number of full hours)

- A future improvement would be to label data by minute, not by hour

Jožef Stefan Institute

Results & discussion #2

- All of the models showed the most misclassified instances with resting. This could be the result of a vague definition of resting in comparison with sleeping, running and most other activities.

- On the other hand, sleeping or smoking are quite well-defined activities where there is little room for subjectivity.

- A prospect for future studies would be to take the most ambiguous or subjective activities and break them down to more defined activities. Although, this would impose greater challenges when collecting data, it could provide more detailed final results.

Confusion matrices

IBk

```
  a   b   c   d   e   f   g   h   i  <-- classified as
345 159 112 160  41   8 102  19  54 |  a = Cleaning.IN
202 327  88 165  47   6  91  28  46 |  b = Cooking.IN
186 144 289 169  84   5  46  26  51 |  c = Playing.IN
147 176 128 243  34  37  94  51  90 |  d = Resting.IN
 88 108  86 116 478   1  44  26  53 |  e = Running.OUT
 27  63  23 108   4 692  28  33  22 |  f = Sleep.IN
149 147 119 176  59   8 250  28  64 |  g = Smoking.IN
165 149 118 143 158  13  39 166  49 |  h = Sports.IN
222 187  75 169  90   4  89  15 149 |  i = Sports.OUT
```

J48

```
  a   b   c   d   e   f   g   h   i  <-- classified as
328 183  62  73  34  26 119 105  70 |  a = Cleaning.IN
122 427  93  43  25  22  83  82 103 |  b = Cooking.IN
140 164 253  85  30   8 147  94  79 |  c = Playing.IN
105 147 109 240  21  36 172  80  90 |  d = Resting.IN
100  64  87  38 512  13  66  59  61 |  e = Running.OUT
 29  15   5  51   8 841  18  31   2 |  f = Sleep.IN
 99 135 107 126  27  33 321  96  56 |  g = Smoking.IN
 85 108 104  81  38  16 111 359  98 |  h = Sports.IN
106 228  98  20  58  14 108  93 275 |  i = Sports.OUT
```

RandomForest

```
  a   b   c   d   e   f   g   h   i  <-- classified as
320 126 106  98  36  27  77  92 118 |  a = Cleaning.IN
104 340 123  58  37  21  85  90 142 |  b = Cooking.IN
101 109 362 116  40  11  79  83  99 |  c = Playing.IN
 91 121 130 267  39  57 107  77 111 |  d = Resting.IN
 71  42  68  51 604  14  30  58  62 |  e = Running.OUT
 17  13   8  60  11 850  14  25   2 |  f = Sleep.IN
 73 107 107 111  45  34 350 106  67 |  g = Smoking.IN
 78  68 106 106  47  20  70 405 100 |  h = Sports.IN
 76 143 107  54  64  14  60  99 383 |  i = Sports.OUT
```

Conclusions

- Combining the data points used in this research with environmental stressors, measured with portable low-cost sensors, could provide detailed results of exposure and intake dose.

- **As low-cost sensors become more widely used and individuals are able to gain access to more information about their living environment, it is crucial for researchers to provide adequate tools to assess and improve accuracy.** A promising step forward would be to reduce the input of individuals and increase the role of machine learning.

Thank you for your attention!

If you have any questions/comments, you can find me at:

✉ rok.novak@ijs.si

in RokNovakSci

R^G Rok Novak

Special thanks to:

Jožef Stefan Institute

ICARUS H2020 project

Slovenian Research agency