Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests

Ruben Brondeel, Bruno Pannier, Basile Chaix

Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests

Ruben Brondeel,[a,b,c] Bruno Pannier,[d] Basile Chaix [a,b]

[a] Inserm, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France
[b] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France
[c] EHESP School of Public Health, Rennes, France
[d] IPC Medical Centre, Paris, France

Corresponding author: Ruben Brondeel, Ruben.Brondeel@gmail.com, +33 1 44 73 86 54, Address: Faculté de médecine Saint-Antoine, 27 rue Chaligny, UMR-S 1136, Nemesis team, 75012 Paris, France

Abstract:

Background: Socioeconomic disparities in active transport have been documented in household travel surveys. However, active transport in these studies was operationalized with self-reported measures, which poorly approximate physical activity. Unfortunately, objective accelerometer data are very expensive to obtain in large-scale travel studies.

Purpose: To benefit from a large sample and objective physical activity data, this study linked a cross-sectional household travel survey with accelerometer data from a small sample to investigate the association between socioeconomic disadvantage and the daily level of transport-related moderate-to-vigorous physical activity (T-MVPA) in an adult population (35-83 years).

Methods: Accelerometer data for participants' trips over 7 days from the RECORD GPS Study (7138 trips, 229 participants) were combined with information on participants' trips over 1 day from the Global Transport Survey (Enquête Globale Transport, EGT) (82084 trips, 21332 participants). Trip-level T-MVPA data from the RECORD sample were used to train a random forests prediction model, enabling the prediction of T- MVPA for each participant's trip from EGT. The associations between socioeconomic indicators and daily T-MVPA were analyzed with negative binomial regression models.

Results: An average time of 18.9 min (95% confidence interval: 18.6-19.2) of T-MVPA was found for these 35-83 year old adults. The education level had a positive association with T-MVPA. Household income had a negative association with T-MVPA, especially for those people without a motorized vehicle.

Conclusions: This study developed a methodology exporting precise sensor- based knowledge to a large survey sample to shed light on population- level socioeconomic disparities in transport-related physical activity.

Keywords

Socially sustainable urban transport; Moderate-to-vigorous physical activity; Transport-related physical activity; Multiple imputation; Random forests; France

Highlights:

- This study analyses transport-related moderate-to-vigorous physical activity (T-MVPA)
- To analyze socioeconomic disparities in T-MVPA, large datasets are needed
- Accelerometer data for large-scale travel studies are very expensive
- Precise accelerometer measures are predicted for a large-scale transport survey
- The education level had a positive association with transport-related MVPA
- The household income had a negative association with transport-related MVPA

1. Introduction

Physical activity is known to be protective for various health outcomes, such as obesity, cardiovascular health problems, depression, and certain cancers (1, 2). The World Health Organization recommends 150 minutes of moderate-to-vigorous physical activity (MVPA) per week for 18 to 64 year old people (3), while the French recommendation is currently of 30 minutes of MVPA per day (4). Transport-related physical activity is an important source of everyday physical activity (5-7), and therefore an important target for health prevention authorities to encourage populations to reach the recommended levels of physical activity.

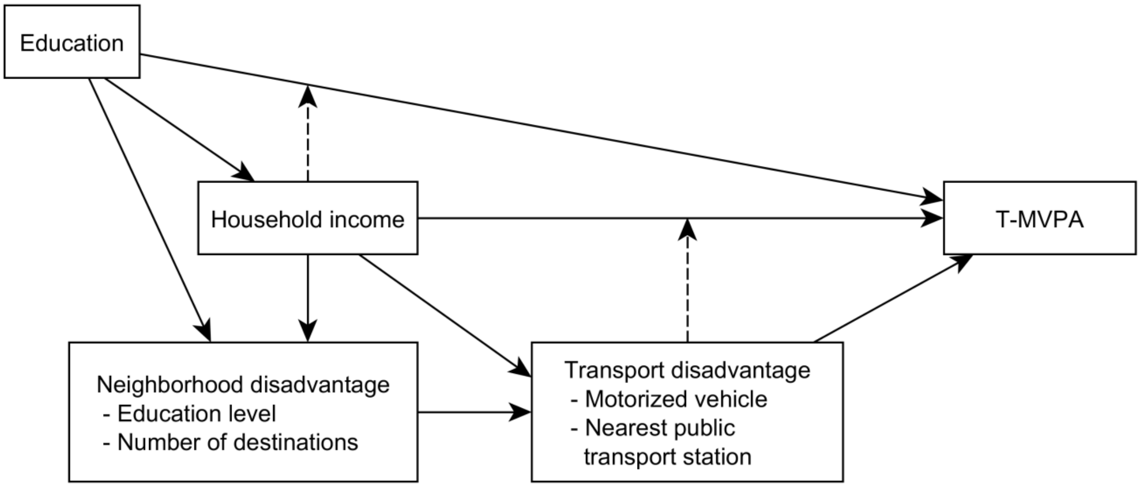
Socioeconomic status leads to disparities in transport-related physical activity (8). For example, a higher personal level of education has been associated with more minutes of walking for transport (9), more trips with active transport modes (9, 10), and more cycling trips (11). In contrast of the finding that higher levels of education are positively associated with active transport, higher income has been associated with fewer minutes of walking and less frequent trips with active modes (9). These results are based on large-scale survey data, as large samples are needed to investigate social inequalities. However, surveys provide only self-reported measures of transport-related physical activity, thus imprecise measures of physical activity: e.g. the ‘usual transportation mode’ or the approximate ‘number of minutes or trips with active transport modes’. These measures are subject to measurement error because people only imprecisely know the start and end times of trips and because they ignore the inactive time during trips with active transportation modes and the physically active time during trips with ‘non-active’ transportation modes (12).

Numerous studies have relied on accelerometers to derive objective measures of physical activity (12, 13). However, studies were less successful in linking transport behavior with physical activity because identifying trips with their exact start and end times is required to perform this linkage (5, 14, 15). Unfortunately, study designs including trip recognition and accelerometer data collection often result in datasets with very precise measures but with limited sample sizes.

An alternative way to measure physical activity for a large number of participants is to rely on a large survey sample and then to estimate the intensity of physical activity based on previously established knowledge. The compendium of Ainsworth (16) enables this by providing an estimated physical activity level in ‘metabolic equivalent of task’ (MET) per minute for numerous activities. The researcher has to determine which category of the compendium relates best to each trip, given the transportation mode, duration of the trip, and intensity of use of certain active modes. However, despite the usefulness of the compendium, the accuracy of its predictions can be criticized. The measures in the compendium are based on findings in very restricted settings (mostly laboratories) (16), and are not adaptable to the characteristics of trips in a specific city or country. Therefore, they may not reflect free-living physical activity in a specific study context.

In this study we present and apply a method that makes predictions for trips reported in a household travel survey based on the data from a GPS and accelerometer data collection conducted in the same geographical context (the Paris Ile-de-France region). The prediction of transport-related physical activity for the trips in the travel survey was based on a random forests model, which enabled us to use a high number of variables to improve the prediction. As a result of this innovative approach, the present study is the first analysis of the effect of socioeconomic status on transport-related MVPA (T-MVPA) in a large and representative dataset of 35 to 83 year old adults (n = 20730).

Figure 1. Directed acyclic graph for the associations between socioeconomic indicators, neighborhood disadvantage, transport disadvantage, and transport-related moderate-to-vigorous physical activity (T-MVPA).



The model studied in this paper is graphically presented (Fig 1) in a directed acyclic graph (DAG). A similar model has been recently tested by Rachele and colleagues (17), describing the relations between educational level, occupational status, household income, neighborhood disadvantage and the most frequently used transportation mode. In our study, the model was applied to T-MVPA instead of the self-reported 'most frequently used transportation mode' and interaction terms were added compared to this previous work, as indicated in the DAG by the dotted arrows (using arrow to arrow notation as suggested by Weinberg (18)). In this model, education and household income were examined separately instead of a single socioeconomic status variable, since these two dimensions had an opposite effect on walking for transport in previous studies (9, 19). The hypothesized interactions are based on findings of social exclusion from transport research. Socioeconomic disadvantage and transport disadvantage (e.g., spatial accessibility to public transport, ownership of a car, or walkability of streets) were found to interact and together amplify social inequalities in the number of trips per individual (20).

This study aimed to investigate the associations between socioeconomic disadvantage, transport disadvantage, and transport-related physical activity for older adults (35 – 83 years old). It expands previous literature by relying on a precise measure of transport-related physical activity and by exploring interactions between various forms of disadvantage. It also describes a novel methodology combining the strengths of a large population dataset with precise sensor-based data (data integration approaches) that advances the field and can be applied to various research questions.

2. Methods

2.1. The Global Transport Survey

The Global Transport Survey (‘Enquête Global Transport’, EGT) is a household travel survey conducted every 10 years in Île-de-France, the French capital region. The main purpose of the survey is to inform local authorities and transport planners on the mobility and transport use in Île-de-France. The latest EGT-survey was conducted in 2010 by two French transport institutions: the Ile-de-France Transport Authority (STIF) and the Regional and Interdepartmental Direction for Equipment and Planning (DRIEA). During face-to-face interviews with members of randomly selected households, data were collected for all the trips made during the day before the interview. We selected participants between 35 and 83 years old for the present study, yielding 82084 trips made by 21332 people. Limiting the EGT-dataset to the people within this age range prevented interpolations of physical activity outside of the age range of the RECORD Study.

2.2. The RECORD GPS Study

As previously described in detail (5, 21), the participants in the RECORD Study (Residential Environment and CORonary heart Disease) were recruited during preventive health checkups in 2007–2008, and were born in 1928–1978. Every participant residing in 112 pre-selected municipalities of the Ile-de-France Paris region at baseline from the administrative files of the IPC Medical Center was invited at the health center (22, 23). The selected municipalities of the region Ile-de-France included a broad range of municipalities in median household income. In the second wave of the study (2011-2012) (24-27), 410 participants were invited to enter the RECORD GPS Study (5). Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (The Actigraph) on the right hip with a dedicated elastic belt, for the recruitment day and 7 additional days, all day long from wake up to bedtime. The participants had to fill out a travel diary by reporting their activity places over the 7-8 days, each time with arrival and departure times. The GPS data were collected every 5 seconds. After linear interpolation of the missing data, the GPS data were analyzed with an algorithm (ArcGIS Python script) that identified all of the activity locations of the participants (any activity at a stationary location) from the accumulation of GPS points over 7 days (28). Based on these outputs of the algorithm, the Mobility Web Mapping application was then used to visualize the activity and transport patterns on a map per participant per day. The Mobility Web Mapping application was designed by the University of Montreal. The application was used to survey the participants on the activity performed at each visited location and on the modes used in each trip. The survey operator could report activity locations and trips undetected by the algorithm and could modify/remove detected visits to locations that were inaccurate or incorrect. This procedure resulted in the identification of 7138 trips for 229 participants. Written informed consent was obtained from all participants. The RECORD GPS Study was approved by the French Data Protection Authority.

Table 1: Educational level and household income of the EGT sample and the RECORD GPS sample

	EGT [a]	RECORD
Educational level		
No diploma of secondary education (%)	40	28
Diploma of secondary education or lower tertiary education [b] (%)	26	30
Diploma of higher tertiary education [c] (%)	33	42
Household income (mean)	3,377	4,393
Sample size	21,332	229

[a] EGT: ‘Enquête globale transport’; [b] Lower tertiary education: two years or less of University education; [c] Higher tertiary education: three years or more of University education

Participants in EGT had a considerable lower education and had a lower household income than the participants in RECORD (see Table 1). Supplementary material S1 provides a comparison of these demographic characteristics between the RECORD sample, the EGT sample and the background population (35 to 83 year old people in Ile-de-France). This comparison supports the hypothesis that the EGT sample represents the background population better than the RECORD sample. The EGT sample included more women, more young people and less people from the inner city.

2.3. Measures

All the dependent and independent variables used in the study are summarized in Table 2.

Table 2: Overview of the variables used in the negative binomial regression model (NB), the multiple imputation model (MI) and the random forests prediction model (RF)

	NB	MI	RF
T-MVPA			
Daily minutes of T-MVPA [a]	X	X	
Minutes of T-MVPA per trip [a]			X
Socioeconomic disadvantage			
Household income [b]	X	X	X
Personal education level [b]	X	X	X
Transport disadvantage			
Street network distance to nearest public transport station from residence [c]	X	X	X
Street network distance to nearest train station [c]			X
Street network distance to nearest metro station [c]			X
Street network distance to nearest tram station [c]			X
Street network distance to nearest bus station [c]			X
A motorized vehicle available in the household [b]	X	X	X
A car available in the household [b]			X
A motorbike available in the household [b]			X
In possession of a public transport pass [b]		X	X
Other personal variables			
Age [b]	X	X	X
Gender [b]	X	X	X
Work situation (employed, unemployed, retired, other) [b]	X	X	X
Other residential neighborhood characteristics			
Educational level in the residential neighborhood [d]	X	X	X
Number of destinations in the residential neighborhood [d]	X	X	X
Number of intersections in the area [d]		X	X
Area size of parks in the area [d]		X	X
Population density in the area [d]		X	X
Address located in Paris, or in the counties adjacent to the city center, or in the counties non-adjacent to the city center		X	X
Personal daily transport behavior			
Minutes in transport per day [e]		X	
Minutes in transport walking per day [e]		X	
Minutes in transport by bike per day [e]		X	
Minutes in transport by private motorized vehicle per day [e]		X	
Minutes in public transport per day [e]		X	
Number of trips per day [e]		X	
Number of trips by walking per day [e]		X	
Number of trips by bike per day [e]		X	
Number of trips by private motorized vehicle per day [e]		X	
Number of trips by public transport day [e]		X	
Trip characteristics			
Transportation mode [f]			X
Duration of the trip in minutes [f]			X

Time of the day at departure [f]		X
Day of the week at departure [f]		X
Rush hour or not at departure: from 8am to 11am and from 4pm to 7pm [f]		X
Straight-line distance from departure address to arrival address [f]		X
Speed based on duration and straight-line distance [f]		X
Trip departure and arrival location characteristics (2 separate set of variables)		
Distance to nearest train station [c]		X
Distance to nearest metro station [c]		X
Distance to nearest tram station [c]		X
Distance to nearest bus station [c]		X
Distance to nearest public transport station [c]		X
Educational level in the area [d]		X
Number of intersections in the area [d]		X
Number of destinations in the area [d]		X
Area size of parks in the area [d]		X
Population density in the area [d]		X
Address located in the city center or not (i.e., in Paris as opposed the other parts of Ile-de-France Region)		X
Day of the EGT mobility survey: week or weekend	X	

[a] Accelerometry information in RECORD or predicted time in EGT; [b] RECORD and EGT questionnaire; [c] Shortest street network distance determined with ArcGIS from the residence or from the departure/arrival of each trip geocoded at the address level in RECORD or at the center of a 100 m square in EGT; [d] The area around the residence or departure or arrival point of each trip was defined with ArcGIS as a 1 km buffer following the street network, and information was aggregated at the level of this buffer; [e] Information from the mobility survey in EGT; [f] Information from the mobility survey in RECORD and in EGT; T-MVPA: transport-related moderate-to-vigorous physical activity.

From the raw accelerometer data, the counts per minute were extracted in ActiLife 5.1. No missing data was allowed within a trip or all data were considered to be missing. There was no minimal wear time per day required. A minute of MVPA was defined as a minute during which a vector magnitude higher than 2690 (29) was recorded, based on the tri-axial GT3X+ accelerometer data in the RECORD GPS study. Accelerometers worn at the hip underestimate physical activity during biking trips. Therefore, all minutes during biking trips were considered as minutes of T-MVPA. This and other limitations of this measure are discussed in the Discussion section.

The following variables were defined both in the RECORD GPS and in the EGT databases (in addition to age and gender). Self-reported household income was coded as a continuous variable. Three educational levels were considered: ‘no diploma of secondary education’, ‘diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education'. Working situation was categorized as employed, unemployed, retired, or other. Participants indicated whether a bike, a motorbike, a car, a motorized vehicle (the combination of the two previous ones) was available in their household. They indicated whether they had a public transport pass. The distance to the nearest public transport station was the distance from the residence to the nearest bus, tram, metro, or train station following the street network. Residential neighborhoods were defined as 1 km buffers around the residence following the street network; corresponding to a 10-to-15 minute walk that reflects the local resources easily accessible within a 'walkable' distance (21, 30-34). The information needed for alternative definitions such as the perceived neighborhood (35) or the activity space (36) was not available. The neighborhood educational level was the percentage of residents with a higher University degree (2010 Census of the National Institute of Statistics and Economic Studies (INSEE)) with census participants geocoded at the building level. The number of destinations in the residential neighborhood was the total number of services of different types (shops, administrative services, leisure facilities, etc.) from the 2011 Permanent Facilities Database of INSEE. We also calculated the number of street intersections (National Geographic Institute data), the area size of parks (Ile-de-France Urbanization Institute), and the population density (2010 Census) in each neighborhood. All these contextual variables were also calculated at the departure and arrival of each trip (see Table 2). ArcGIS (v10.3) automated using Python (v2.7) was used for the geographical analyses.

Based on the RECORD and EGT mobility surveys, the following variables were determined at the trip level: transportation mode, trip duration, time and day of the trip, distance covered and speed.

2.4. Statistical analysis

An overview of the dependent and independent variables in the prediction model, the multiple imputation model and the main regression model is provided in Table 2.

The RECORD GPS data were used to train a random forests prediction model for T-MVPA (see explanatory variables in Table 2) with 1000 trees and a random selection of 16 variables at each knot. The random forests model was grown with the 'randomForest' package (37) in R. Based on the prediction model and on the comparable prediction variables in EGT, we predicted the number of minutes of T-MVPA for each trip in the EGT dataset. The predicted values were summed up per day, resulting in a daily time of T-MVPA in minutes per person.

The associations between the disadvantage variables and the predicted T-MVPA time were analyzed with a negative binomial regression model using the 'MASS' package in R (38). The time variable could be considered as continuous and analyzed with a regular linear regression. However, given the left-censored distribution of the variable (i.e. 0 as the absolute minimum and many observation equal 0 or close to 0), we preferred the negative binomial regression that is adapted to count variables with overdispersion (a high variance compared to the mean). There were missing values on 8 independent variables for 24 % of the respondents, of which 6 % had more than 1 missing value. Therefore, multiple imputations were performed with the 'mice' package in R (39). This method enabled us to analyze the data under the hypothesis that the unobserved values are randomly distributed given the observed data (40). To account for the non-linear and interaction effects in the imputation process, random forests methods were also used for the multiple imputations of explanatory variables in EGT. Five imputation datasets were constructed though an iterative process using 100 trees for every imputed variable at each iteration. One imputed dataset was retained every five iterations (25 iterations overall). The convergence of the imputations was checked with plots of the means and standard deviations over the iterations.

In the analysis of the determinants of T-MVPA, the interaction terms of interest were plotted in graphs based on the coefficients and on the variance-covariance matrix from the regression model. The code for these plots was based on the library 'effects' in R (41), but adapted to the negative binomial regression. The script for all the analyses with R (v3.2.2) (42) can be found in Supplementary material S2.

3. Results

The random forests prediction model for T-MVPA was very accurate, predicting 67% of the variance in T-MVPA in RECORD. The three most important variables in predicting trip-level T-MVPA were transportation mode, distance and duration of the trip (see Supplementary material S2). Applying this model to the EGT trips and summing up the predicted minutes of T-MVPA by day, we found a mean predicted time of T-MVPA of 18.9 minutes (95% confidence interval (CI): 18.6-19.2) per participant per day (interquartile range: 5, 28). The mean T-MVPA times for the levels of education 'no diploma of secondary education', 'diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education' were respectively of 17.5, 18.5, and 21.0 minutes per day (descriptive data, unadjusted). Household income was negatively associated with the daily T-MVPA time (Incidence Risk Ratio = 0.98 for a change in income of 1000€, 95% CI: 0.97-0.99). Regarding transportation disadvantage, participants who had access to a motorized vehicle (i.e., a car or motorbike) in the household had a mean daily T-MVPA time of only 16.8 minutes while their counterparts who had no vehicle had 28.9 minutes of T-MVPA per day. The distance to the nearest public transport station was negatively associated with the daily T-MVPA time (Incidence Risk Ratio = 0.72 for a change in distance of 1km, 95% CI: 0.65-0.78). No difference between men and women was noted. Finally, older people had slightly less daily T-MVPA (Incidence Risk Ratio = 0.98 for a change in age of 10 years, 95% CI: 0.97-1.00).

Table 3: Associations between socioeconomic or transport disadvantage and daily T-MVPA (negative binomial regression)

Predictor	IRR	95% CI
Socioeconomic disadvantage		
Education level		
No diploma of secondary education	1.00	Referent
Diploma of secondary education or lower tertiary education [a]	1.06	1.01, 1.10
Diploma of higher tertiary education [b]	1.12	1.07, 1.17
Household income (/1000 euros)	0.97	0.94, 1.00
Interaction Education –Income		
No secondary education - income	1.00	Referent
Secondary or lower tertiary education - income	0.99	0.96, 1.02
Higher tertiary education - income	1.00	0.98, 1.03
Transport disadvantage		
Motorized vehicle available in household		
No motorized vehicle	1.00	Referent
Motorized vehicle	0.65	0.61, 0.68
Nearest public transport (km)	1.01	0.93, 1.11
Interactions Socioeconomic - Transport		
Motorized vehicle - income		
No motorized vehicle	1.00	Referent
Motorized vehicle	1.02	1.00, 1.05
Nearest public transport - income	0.95	0.90, 1.00
Neighborhood disadvantage		
Educational level	1.26	1.11, 1.43
Number of destinations (/1000)	1.12	1.10, 1.14
Other		
Age (10y)	0.96	0.94, 0.98
Gender		
Female	1.000	Referent
Male	1.02	0.99, 1.05
Work situation		
Employed	1.000	Referent
Unemployed	1.02	0.95, 1.09
Retired	1.09	1.03, 1.14
Other	0.98	0.92, 1.04
(intercept)	24.19	22.71, 25.77

Abbreviations: CI, confidence interval; IRR, incidence rate ratio; MVPA, moderate-to-vigorous physical activity, [a] Lower tertiary education: two years or less of University education; [b] Higher tertiary education: three years or more of University education

The results of the multiple negative binomial regression (Table 3) confirmed the bivariate analyses, while adding nuance by introducing interaction effects. Figure 2 and Figure 3 represent two interaction effects. Household income had a negative association with T-MVPA for all three categories of education level (Figure 2). The interaction effect was statistically significant (Wald-test for pooled regression results (39): $P = 0.041$), but there was no clear gradient in the strength of the association between income and T-MVPA between the different education levels. Furthermore, household income had a negative association with T-MVPA for both those with and without a motorized vehicle available in the household (Figure 3). However, the association was much stronger for those without a motorized vehicle.

The distance to the nearest public transport station had a negative association with T-MVPA for all levels of income. The interaction effect with income was small and does not alter the interpretation of the results. Two interaction effects between education level and transport disadvantage (availability of a motorized vehicle and access to public transport) were tested. Including these into the model did not change the interpretation of the results nor did it improve the model in statistical terms ($P = 0.180$). For the sake of parsimony, these two interaction terms were excluded from the final model.

Figure 2. Moderating effect of household income on the relationship of personal education level to daily minutes of transport-related moderate-to-vigorous physical activity (T-MVPA). Confidence intervals and predicted values are represented for each category of education level for all levels of household income within the studied range.

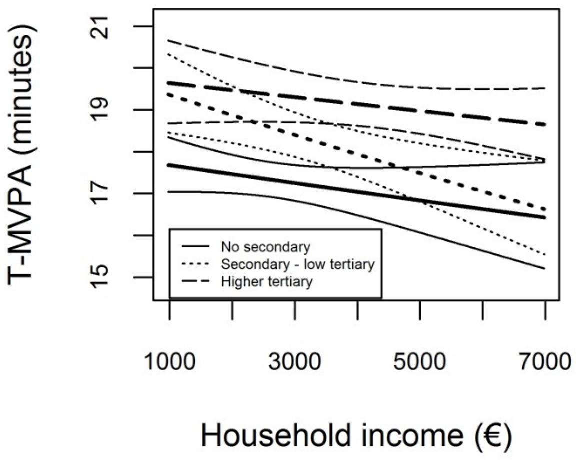
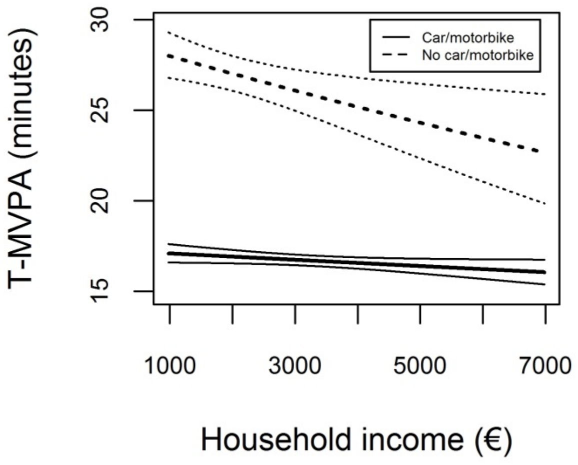


Figure 3. Moderating effect of the availability of a motorized vehicle in the household on the relationship of household income to daily minutes of transport-related moderate-to-vigorous physical activity (T-MVPA). Confidence intervals and predicted values are represented for the availability of a vehicle (yes/no) for all levels of household income within the studied range.



To facilitate the interpretation of the associations between socio-economic factors and T-MVPA, Table 4 provides information on the associations between educational level and household income on one hand and the mean number of trips and the mean duration of trips by transportation mode on the other hand. From these descriptive data, the positive association of educational level with T-MVPA may be attributable to some extent to the number of walking and public transport trips. Higher educated people had more walking and public transport trips. This is attenuated but not completely counterbalanced by the longer duration of walking and public transport trips of lower educated people. The negative association of income with T-MVPA may also be attributable to some extent to the number of walking and public transport trips and to the duration of the walking trips. People with higher income had less and shorter walking trips, and less public transport trips. From the descriptive data, biking trips had little or no impact on both associations.

Table 4: Mean number of trips and mean duration of trips in the EGT sample and in the RECORD GPS sample

	Mean number of trips per person				Mean duration of trips (min)			
	W	B	PM	PT	W	B	PM	PT
Education [a]								
Level 1	1.2	0.0	1.8	0.4	14.2	25.0	22.4	52.8
Level 2	1.2	0.0	2.2	0.5	13.1	24.6	21.9	50.8
Level 3	1.4	0.1	2.0	0.7	12.5	20.1	22.5	45.9
Household income								
Less than 2000 €	1.5	0.0	1.3	0.7	14.0	19.8	22.8	49.8
2000 to 4000 €	1.2	0.0	2.1	0.5	13.3	23.6	22.0	50.4
4000 € or more	1.2	0.1	2.3	0.5	12.2	22.2	22.7	47.1

W: walking; B: biking; PM: private motorized (car/motorbike); PT: public transport; min: minutes; [a] Education: No diploma of secondary education, Diploma of secondary education or lower tertiary education (2 years or less of University education); Diploma of higher tertiary education (three years or more of University education)

4. Discussion

4.1. Main results

Our study suggests that transport-related physical activity is a major source of physical activity for the population in the Ile-de-France region. On average, the participants had 18.9 minutes of MVPA per day. The international recommendation of 30 minutes of MVPA per day (including all sources of physical activity) was attained by 23% of participants through their transport behavior alone.

The model showed a negative association of household income with T-MVPA and a positive relation of educational level with T-MVPA. Understanding the mechanisms underlying these associations is very important to efficiently target subpopulations in physical activity interventions. It has been argued that lower educated people have symbolic and affective predispositions that promote car use over active transport (e.g., car use perceived as a marker of wealth) (10, 43). Instead of psychological explanations, other studies have established a link between lower educational levels and material obstacles to healthy behavior including physical activity (30, 44). These obstacles are situated within diverse domains of the social life: e.g., the residential environment (e.g. walking possibilities) or the workplace (e.g. parking facilities at work) or the local organization of transport (e.g. bus frequency) (45, 46). Further research is needed to fully understand the motivations and obstacles of people with a lower level of education and a high income to participate in active transport, and to confirm the observed patterns of associations in other geographical contexts and other populations such as children going to school or younger adults. However, the results clearly show that education and income should be considered separately when studying transport-related physical activity or mobility in general, instead of using a combined measure of socioeconomic status.

The availability of a motorized vehicle largely moderated the association between household income and T-MVPA. The negative association of household income with T-MVPA was much stronger within the group of people with no motorized vehicle available. This might reflect the influence of the distance from the residence to important places such as work or services. For the higher income groups, this distance is typically shorter than for the lower income groups. So, people with long trips to cover and no accessibility to a motorized vehicle are constrained to use more active transport modes, including public transport.

4.2. Strengths and limitations

Hopefully, technical advances will enable researchers in the future to both assess the trips of study participants and objectively measure physical activity in these trips for large samples of people. Until then, we believe that predicting transport-related physical activity (here T-MVPA) by applying precise knowledge derived from sensor data to large survey datasets has several advantages over the use of approximate self-reported measures (e.g., on the use of active transport) combined with information from a physical activity compendium. Compared to approximate self-reported measures, T-MVPA enables the comparison to the WHO health recommendations; it allows one to take into account the specific intensity of physical activity of active modes in the study territory of interest; and it includes the physical activity during trips with ‘non-active’ modes (e.g., walk to or from a car, use of stairs in public transport). This is especially important in regions with a relatively high use of public transport, such as in the French capital region Île-de-France. Daily T-MVPA is a useful variable from a public health perspective since it encompasses the influences of the transportation mode, the number of trips, and the duration of trips instead of just one of these indicators. Moreover, compared to the use of a compendium, the prediction of T-MVPA is based on sensor data from the same geographical context. Finally, the use of an underlying prediction model enables the use of numerous variables to individualize the physical activity intensity for each participant’s profile. Therefore, it can be expected that the predictions are of much better quality than if standard compendium values were applied to trips, even though a comparative study is needed to examine this.

This study provided a sophisticated model including direct, moderated, and mediated associations between socioeconomic disadvantage and T-MVPA. Especially the moderated associations presented in this study show the need for a conceptual thinking that goes beyond basic associations applied to everyone when investigating social disparities in T-MVPA. Unfortunately, we could not test other variables of the built environment (e.g., the width of sidewalks) than those that were examined, or related to other individual dimensions (time available, behavioral preferences, etc.) to further understand and explain the associations between education, income, and T-MVPA.

Combining two datasets from the same geographical setting – a large-scale survey and a smaller dataset with detailed sensor measures – could be a pragmatic approach to address a large range of research questions where large data collections with detailed measures are too expensive. Given a good prediction model with variables available in both datasets, this method could provide a relatively inexpensive option for research questions where large-scale survey data are necessary (e.g., when investigating population disparities as in our case). Further methodologic work is needed to evaluate different machine learning methods. The random forests method was preferred for this study, since it explained a high percentage of the variation (67%) compared to two other machine learning methods: support vector machines (42%, using the ‘svm’ function in the R package ‘e1071’ (47)) and neural networks (45%, using the ‘mlp’ function in the R package ‘RSNNS’ (48)). Secondly, the random forests method does not rely on parameters of the distribution of the outcome variable. Therefore, it cannot predict values outside the range of the input data, which is particularly important for a left-censored variable (i.e. 0 as a strict minimum value) such as T-MVPA. A limitation of the random forests method is its complexity, making it hard to interpret the relations between the predictive variables and the outcome.

The cut point for MVPA used in this study is not without limitation. The cut point aims to identify body movements that require an energy expenditure of three MET (metabolic equivalent of task) (29). The cut point is not age-specific, whereas research has found that the energy expenditure is higher for older people than younger people when performing the same physical task (49). The cut point will therefore have to be age-specific in future research. Also, the cut point has been established during laboratory tests and might therefore poorly correspond to three METs in free-living conditions.

An important limit to this study is the lack of a total daily MVPA measure (e.g., including leisure physical activity). A lack of transport-related physical activity could be compensated by leisure-time physical activity. And even though this compensation mechanism was documented neither by Hearst (50) for walking time nor by Sahlqvist (7) for self-reported physical activity, more studies in this domain are needed.

Finally, for biking trips, an accelerometer at the hip usually underestimates T-MVPA. Therefore, we had to use an estimate of biking physical activity from the compendium of Ainsworth (16). A drawback of this is that all minutes of biking trips were considered to be physically active, disregarding stops over the way. The impact on the results is probably small with around 6.2% of

T-MVPA obtained from cycling in this population. A slight overestimation of this small share of T-MVPA probably only led to a minor overestimation of the daily T-MVPA. For studies with cycling as the focus, other types of accelerometer devices (such as the VitaMove system used in the RECORD MultiSensor Study) or other ways to carry the accelerometer are recommended.

5. Conclusions

This study is, to our knowledge, the first to use a large dataset to estimate the association between socioeconomic disadvantage and T-MVPA. It gives insights on the relationships between socioeconomic disadvantage and daily transport-related physical activity, which is a relatively large part of the daily physical activity of the adult population in the Ile-de-France region. An important finding for future interventions on active transport is that both the expected positive association with education and a negative association with income were document. More research is needed to understand the exact motivations and obstacles leading to social disparities in transport-related physical activity.

Acknowledgements

The RECORD GPS Study was supported by INPES (National Institute for Prevention and Health Education); the Ministry of Ecology (DGITM); CEREMA (Centre for the Study of and Expertise on Risks, the Environment, Mobility, and Urbanism); ARS (Health Regional Agency) of Ile-de-France; STIF (Ile-de-France Transport Authority); the Ile-de-France Regional Council; RATP (Paris Public Transport Operator); and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning). The authors thank the following partners from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Babaa, Laurent Jardinier, Tristan Guilloux, Mélani Alberto, Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré.

References

1. Wanner M, Gotschi T, Martin-Diener E, Kahlmeier S, Martin BW. Active transport, physical activity, and body weight in adults: a systematic review. Am J Prev Med. 2012;42(5):493-502.

2. de Nazelle A, Nieuwenhuijsen MJ, Anto JM, Brauer M, Briggs D, Braun-Fahrlander C, et al. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. Environ Int. 2011;37(4):766-77.

3. World Health Organisation. Factsheet on physical activity 2015 [Available from: http://www.who.int/mediacentre/factsheets/fs385/en/.

4. Programme National Nutrition Santé. Manger Bouger - Que veut dire bouger? 2015 [Available from: http://www.mangerbouger.fr/bouger-plus/que-veut-dire-bouger.html.

5. Chaix B, Kestens Y, Duncan S, Merrien C, Thierry B, Pannier B, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int J Behav Nutr Phy. 2014;11(1):124.

6. Besser LM, Dannenberg AL. Walking to public transit: steps to help meet physical activity recommendations. Am J Prev Med. 2005;29(4):273-80.

7. Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. Prev Med. 2012;55(3):206-11.

8. Beenackers MA, Kamphuis CB, Giskes K, Brug J, Kunst AE, Burdorf A, et al. Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. Int J Behav Nutr Phys Act. 2012;9:116.

9. Cerin E, Leslie E, Owen N. Explaining socio-economic status differences in walking for transport: an ecological analysis of individual, social and environmental factors. Soc Sci Med. 2009;68(6):1013-20.

10. Scheepers E, Wendel-Vos W, van Kempen E, Panis LI, Maas J, Stipdonk H, et al. Personal and environmental characteristics associated with choice of active transport modes versus car use for different trip purposes of trips up to 7.5 kilometers in The Netherlands. PLoS One. 2013;8(9):e73105.

11. Carse A, Goodman A, Mackett RL, Panter J, Ogilvie D. The factors influencing car use in a cycle-friendly city: the case of Cambridge. J Transp Geogr. 2013;28(100):67-74.

12. Steene-Johannessen J, Anderssen SA, van der Ploeg HP, Hendriksen IJ, Donnelly AE, Brage S, et al. Are Self-Report Measures Able to Define Individuals as Physically Active or Inactive? Med Sci Sport Exer. 2016;48(2):235-44.

13. Wijndaele K, Westgate K, Stephens SK, Blair SN, Bull FC, Chastin SF, et al. Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. Med Sci Sport Exer. 2015;47(10):2129-39.

14. Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transport Res C-Emer. 2009;17(3):285-97.

15. Brondeel R, Pannier B, Chaix B. Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. Med Sci Sport Exer. 2015;47(12):2669-75.

16. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR, Jr., Tudor-Locke C, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sport Exer. 2011;43(8):1575-81.

17. Rachele JN, Kavanagh AM, Badland H, Giles-Corti B, Washington S, Turrell G. Associations between individual socioeconomic position, neighbourhood disadvantage and transport mode: baseline results from the HABITAT multilevel study. J Epidemiol Community Health. 2015.

18. Weinberg CR. Can DAGs clarify effect modification? Epidemiology. 2007;18(5):569-72.

19. Turrell G, Hewitt B, Haynes M, Nathan A, Giles-Corti B. Change in walking for transport: a longitudinal study of the influence of neighbourhood disadvantage and individual-level socioeconomic position in mid-aged adults. Int J Behav Nutr Phys Act. 2014;11:151.

20. Lucas K. Transport and social exclusion: Where are we now? Transp Policy. 2012;20:105-13.

21. Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: The effect of spatial accessibility to healthcare services. Health Place. 2014;30C:127-33.

22. Chaix B, Bean K, Daniel M, Zenk SN, Kestens Y, Charreire H, et al. Associations of supermarket characteristics with weight status and body fat: a multilevel analysis of individuals within supermarkets (RECORD Study). PLoS One. 2012;7(3):e32908.

23. Van Hulst A, Thomas F, Barnett TA, Kestens Y, Gauvin L, Pannier B, et al. A typology of neighborhoods and blood pressure in the RECORD Cohort Study. J Hypertens. 2012;30(7):1336-46.

24. Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, et al. Cohort Profile: Residential and non-residential environments, individual activity spaces and

cardiovascular risk factors and diseases--The RECORD Cohort Study. Int J Epidemiol. 2012;41(5):1283-92.

25. Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. Am J Prev Med. 2012;43(4):440-50.

26. Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in the associations between multiple environmental features and body weight and abdominal fat: using matching techniques to assess whether the associations are separable. Am J Epidemiol. 2012;175(11):1152-62.

27. Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). Soc Sci Med. 2014;119:64-73.

28. Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. Int J Health Geogr. 2013;12(14).

29. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport. 2011;14(5):411-6.

30. Chaix B, Simon C, Charreire H, Thomas F, Kestens Y, Karusisi N, et al. The environmental correlates of overall and neighborhood based recreational walking (a cross-sectional analysis of the RECORD Study). Int J Behav Nutr Phy. 2014;11(1):20.

31. Frank LD, Schmid TL, Sallis JF, Chapman J, Saelens BE. Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. Am J Prev Med. 2005;28(2 Suppl 2):117-25.

32. Karusisi N, Thomas F, Meline J, Brondeel R, Chaix B. Environmental conditions around itineraries to destinations as correlates of walking for transportation among adults: the RECORD cohort study. PLoS One. 2014;9(5):e88929.

33. Troped PJ, Wilson JS, Matthews CE, Cromley EK, Melly SJ. The built environment and location-based physical activity. Am J Prev Med. 2010;38(4):429-38.

34. Villanueva K, Knuiman M, Nathan A, Giles-Corti B, Christian H, Foster S, et al. The impact of neighborhood walkability on walking: does it differ across adult life stage and does neighborhood buffer size matter? Health Place. 2014;25:43-6.

35. Vallée J, Le Roux G, Chaix B, Kestens Y, Chauvin P. The 'constant size neighbourhood trap' in accessibility and health studies. Urban Studies. 2015;52(2):338-57.

36. Matthews SA, Yang TC. Spatial Polygamy and Contextual Exposures (SPACEs): Promoting Activity Space Approaches in Research on Place and Health. Am Behav Sci. 2013;57(8):1057-81.

37. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

38. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth Edition ed. New York: Springer; 2002.

39. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1-67.

40. Little RJA, Rubin DB. The Analysis of Social Science Data with Missing Values. Sociological Methods & Research. 1989;18(2-3):292-326.

41. Fox J. Effect Displays in R for Generalised Linear Models. J Stat Softw. 2003;8(15):1-27.

42. R Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing V, Austria, editor. 2014.

43. Beirão G, Sarsfield Cabral JA. Understanding attitudes towards public transport and private car: A qualitative study. Transp Policy. 2007;14(6):478-89.

44. Brunello G, Fort M, Schneeweis N, Winter-Ebmer R. The Causal Effect of Education on Health: What Is the Role of Health Behaviors? Health Econ. 2016;25(3):314-36.

45. Delbosc A, Currie G. Transport problems that matter – social and psychological links to transport disadvantage. J Transp Geogr. 2011;19(1):170-8.

46. Dalton AM, Jones AP, Panter JR, Ogilvie D. Neighbourhood, Route and Workplace-Related Environmental Characteristics Predict Adults' Mode of Travel to Work. PLoS One. 2013;8(6):e67575.

47. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. 2015.

48. Bergmeir C, Benitez JM. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. J Stat Softw. 2012;46(7):1-26.

49. Hall KS, Howe CA, Rana SR, Martin CL, Morey MC. METs and accelerometry of walking in older adults: standard versus measured energy cost. Med Sci Sports Exerc. 2013;45(3):574-82.

50. Hearst MO, Sirard JR, Forsyth A, Parker ED, Klein EG, Green CG, et al. The relationship of area-level sociodemographic characteristics, household composition and individual-level socioeconomic status on walking behavior among adults. Transp Res Part A Policy Pract. 2013;50:149-57.

Supplementary material

Supplementary material S1: Overview of the demographic characteristics of the background population (35-83 year old people in Ile-de-France), in the EGT sample and in the RECORD GPS sample

	I-d-F [a] (%)	EGT [b] (%)	RECORD (%)
Gender			
Female (%)	52	53	37
Male (%)	48	47	63
Age [c]			
35-44 years	30	32	16
45-59 years	39	37	37
60-74 years	24	25	41
74-83 years	7	6	7
Location of residence [d]			
Inner city (Paris)	19	14	27
First crown of counties around Paris	37	36	42
Second crown of counties around Paris	44	51	31
Population / sample size	5,887,647	21,332	229

[a] I-d-F: 2012 Census data from Ile-de-France, the French capital region; [b] EGT: 'Enquete globale transport'; [c] The data for the age groups 35-44 and 74-83 were not available in the population statistics. The percentages for these categories are based on the assumption that the distribution within the broader category is uniform; [d] The categorization of urbanicity is based on an official administrative subdivision of the Ile-de-France region.

Supplementary material S2: R-scripts

```
library(data.table)
library(randomForest)
library(mice)

######################################################
# A Construct prediction model MVPA based on RECORD data
######################################################
path <- "~/.../data/"
rec <- data.table(read.csv(paste0(path, "1. RECORD.csv")))

# 1. Impute missing values in RECORD dataset
# imputations are based on a Random Forest multiple imputation (1 iteration)

# 1.1 order variables in number of missing values.
# this will help the efficiency of the imputation process
seq <- dimnames(md.pattern(rec[,4:ncol(rec), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c("trip_code", "depcom_res", "dciris_res"), seq)
rec <- rec[, seq, with=FALSE]

# 1.2 Use mice() with the maximum number of iterations maxit set to zero.
# This is a fast way to create the mids object called ini
# containing the default settings.
# (Van Buuren S, Groothuis-Oudshoorn K.
# mice: Multivariate Imputation by Chained Equations in R.
# Journal of Statistical Software. 2011;45(3):1-67.)
rec.ini <- copy(rec)

rec.ini[,':=' (trip_code='1', depcom_res = 1, dciris_res= 1)]
ini <- mice(rec.ini, max=0, meth='rf')

meth <- ini$meth
pred <- ini$pred
vis <- ini$vis

# 1.3 use these setting
# Method (meth, here random forest), predictors per variable imputed (pred) and
# visiting sequence (vis)
mi.rf <- mice(rec, m=1, maxit = 5, pred=pred, meth=meth, vis=vis)

# 1.4 creating a dataset with all missings imputed
rec.nomiss <- complete(mi.rf)

# 2 MVPA prediction model on RECORD data
```

```
form.mv <- formula(mvpa_ep1m ~ mode_trans1 + duration_mn +
                   time_of_day + day_trip + rush_hour +
                   age + homme + dist_ld + speed_ld + rvnu +
                   emploi_sim + nivetude_sim +
                   dist_train_dep + dist_metro_dep + dist_tram_dep + dist_bus_dep +
                   dist_train_arr + dist_metro_arr + dist_tram_arr + dist_bus_arr +
                   dist_train_res + dist_metro_res + dist_tram_res + dist_bus_res +
                   dist_pt_res + dist_pt_dep + dist_pt_arr +
                   educ_res + educ_dep + educ_arr +
                   intersec_res + intersec_dep + intersec_arr +
                   dest_res + dest_dep + dest_arr +
                   park_res + park_dep + park_arr +
                   pdens_res + pdens_dep + pdens_arr +
                   res_cour + dep_cour + arr_cour +
                   pos.voiture + pos.moto + pos.TC + pos.motorized)

fit.mvp <- randomForest(form.mv, data=rec.nomiss, ntree = 1000)


# list of 15 most important variables

a <- data.frame(importance(fit.mvp))
a$Variables <- rownames(a); rownames(a) <- NULL
a[order(a$IncNodePurity, decreasing=TRUE),c('Variables', 'IncNodePurity')][1:15,]

#         Variables IncNodePurity
# 1     mode_trans1    158005.805
# 2     duration_mn    113574.763
# 8         dist_ld     54074.314
# 9        speed_ld     48475.942
# 18 dist_metro_arr     12746.681
# 19  dist_tram_arr     12178.505
# 4        day_trip     10863.437
# 39       park_arr      9102.616
# 14 dist_metro_dep      8925.460
# 38       park_dep      8772.124
# 13 dist_train_dep      8584.357
# 41      pdens_dep      8534.186
# 33   intersec_arr      8323.625
# 32   intersec_dep      8078.804
# 15  dist_tram_dep      8004.453

# to visualize importance of variables
varImpPlot(fit.mvp)

##########################################################
# B Prediction of MVPA for EGT trips
```

```
##################################################################
path <- "~/.../data/"
egt <- data.table(read.csv(paste(path, "2. EGT.csv", sep="")))

# 1 Imputation of missing values in EGT datasets
#   the imputation will enable MVPA predictions for all trips
#   The imputations are based on predictive mean matching models

# 1.1 Ordering the variables on the amount of missing values
#     while making sure id variables won't be used in the imputation
seq <- dimnames(md.pattern(egt[,5:ncol(egt), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c("trip_code", "resc", "depcom_res", "dciris_res"), seq)
egt <- egt[, seq, with=FALSE]
egt.ini <- copy(egt)
egt.ini[,':=' (trip_code='1', resc = '1', depcom_res = 1, dciris_res= 1)]

# 1.2 Use mice() with the maximum number of iterations maxit set to zero.
#     This is a fast way to create the mids object called ini
#     containing the default settings.

ini <- mice(egt.ini, max=0)
meth <- ini$meth
pred <- ini$pred # since in egt.ini the id variables are constant, the pred is already==0
vis <- ini$vis

# 1.3 Actual imputation of EGT dataset
mi.data <- mice(egt, m=1, maxit = 5, pred=pred, meth=meth, vis=vis)
egt.nomiss <- complete(mi.data)


# 2. Prediction of MVPA for each EGT trip
#   Using the new egt.nomiss dataset and
#   MVPA-prediction model on RECORD data
#   Note: the original not-imputed EGT dataset is used after this step
#   egt.nomiss is only used for these predictions
egt[, pred.mvpa := predict(fit.mvp, egt.nomiss)]


# 3. Some variables for the regression analysis
# 3.1 Variable Weekend-Weekday
tmp <- data.table(day_trip=c('1. Monday', '2. Tuesday', '3. Wednesday', '4. Thursday', '5. Friday',
'6. Saturday', '7. Sunday'),
                weekday=as.factor(c(rep('1. weekday', 5), rep('2. weekend',2))))
egt <- merge(egt, tmp, by='day_trip', all.x=T)

# 3.2 Creating an id variable for the person
```

```
a <- unlist(strsplit(as.character(egt$trip_code), '_'))
men <- a[seq(1,length(a), 3)]
per <- a[seq(2,length(a), 3)]
egt$person <- paste(men, per, sep='_')

####################################################
# C Construction of day-level EGT dataset
####################################################
# 1. Construction of day level variables
# 1.1 MVPA per day, minutes in transport per day and number of trips
setkey(egt, person)
egt[,V1 := 1]

var1 <- c('pred.mvpa',  'duration_mn', 'V1') #'pred.mf.mvpa',
var2 <- c('mvpa.day',  'min.day', 'nb_trips') #'mvpa.mf.day',
egt[, var2 := lapply(.SD, sum, na.rm=TRUE), by=person, .SDcols=var1, with=FALSE]


# 1.2 Day-level variables per type of transportation mode
# 1.2.1 MVPA (so how much each person profits of each transportation mode in terms of
MVPA)

setkey(egt, person, mode_trans1)
mvpa_by_mt <- egt[, sum(pred.mvpa),by=list(person,mode_trans1)]

setkey(mvpa_by_mt, person, mode_trans1)
out <- mvpa_by_mt[CJ(unique(person), unique(mode_trans1))][, as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep=''),
        paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"), sep=''))
var <- paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"), sep='')
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]

egt <- merge(egt, out, by='person')

# 1.2.2 Minutes in transport

setkey(egt, person, mode_trans1)
min_by_mt <- egt[, sum(duration_mn),by=list(person,mode_trans1)]

setkey(min_by_mt, person, mode_trans1)
out <- min_by_mt[CJ(unique(person), unique(mode_trans1))][, as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep=''),
        paste("MIN_", c('NA', "walking", "biking", "PM", "PT"), sep=''))
var <- paste("MIN_", c('NA', "walking", "biking", "PM", "PT"), sep='')
```

```
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]

egt <- merge(egt, out, by='person')

# 1.2.3 Number of trips

setkey(egt, person, mode_trans1)
nb_by_mt <- egt[, sum(V1), by=list(person,mode_trans1)]

setkey(nb_by_mt, person, mode_trans1)
out <- nb_by_mt[CJ(unique(person), unique(mode_trans1))][, as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep="),
        paste("nb_", c('NA', "walking", "biking", "PM", "PT"), sep="))
var <- paste("nb_", c('NA', "walking", "biking", "PM", "PT"), sep=")
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]
egt[,V1 := NULL]
egt <- merge(egt, out, by='person')

# 2. Add people with no trips
#   2066 people were not in the trip-dataset,
#   because they reported no trips at all during the day of observation

egtnt <- data.table(read.csv(paste(path, '2. EGT no trips.csv', sep="")))

# 2.1 Create the variables in egtnt that were created before in EGT dataset
varnt <- names(egtnt)[which(names(egtnt) %in% names(egt))]
egtnt2 <- egtnt[,varnt, with=FALSE]

# 2.2 Set these variables to 0
#   (e.g. no transport-related MVPA observed for these people)
egtnt2[, var2:= 0, with=FALSE]
egtnt2[, paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"), sep="):= 0, with=FALSE]
egtnt2[, paste("MIN_", c('NA', "walking", "biking", "PM", "PT"), sep="):= 0, with=FALSE]
egtnt2[, paste("nb_", c('NA', "walking", "biking", "PM", "PT"), sep="):= 0, with=FALSE]
egtnt2[, c('min.day', 'nb_trips'):= 0, with=FALSE]

# 2.3 Merge EGT dataset with EGT no trips dataset
egt <- rbindlist(list(egt, egtnt2), use.names=TRUE, fill=TRUE)


# 3. aggregate to day level
egt.day <- unique(setkey(egt, person), by='person')
```

```
########################################################
# D recode some variables for the analysis
########################################################

# 1 Round mvpa variable to the minute.
# This is necessary for count regression
egt[, mvpa.day.int := round(mvpa.day)]

# 2 Centralize variables for easier interpretable interaction effects
# and divide variables by 1000 to get an interpretable scale (e.g. km)

egt[, rvnu.1000 := (rvnu - mean(rvnu, na.rm=TRUE))/1000]
egt[, age.10 := ( age - mean(age, na.rm=TRUE))/10]
egt[, intersec_res.1000 := (intersec_res - mean(intersec_res, na.rm=TRUE))/1000]

egt[dist_pt_res>1000, dist_pt_res := 1000]
egt[, dist_pt_res.1000 := (dist_pt_res - mean(dist_pt_res, na.rm=TRUE))/1000]

egt[, educ_res.m := (educ_res  - mean(educ_res, na.rm=TRUE ))]
egt[, dest_res.1000 := (dest_res - mean(dest_res, na.rm=TRUE))/1000]



########################################################
# E Multiple imputation of EGT day-level dataset
########################################################
# 1 : imputation of missing values to have a MVPA prediction for all trips

# 1.1 ordering the variables on the amount of missing values
# while making sure id variables won't be used in the imputation
seq <- dimnames(md.pattern(egt[,6:ncol(egt), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c('person', 'resc', 'depcom_res', 'dciris_res', 'over'), seq)
egt <- egt[, seq, with=FALSE]

# 1.2 Use mice() with the maximum number of iterations maxit set to zero.
# This is a fast way to create the mids object called ini
# containing the default settings.
egt.ini <- copy(egt)

egt.ini[,':=' (person='1', resc = '1',  depcom_res = 1, dciris_res= 1, over = 1)]
ini <- mice(egt.ini, max=0, meth='rf')

meth <- ini$meth
meth[c("person", "resc", "depcom_res", "dciris_res",
      "mvpa.day.int", "age.10", "homme",  "res_cour",  "pos.motorized",
      "weekday",  "min.day",  "MIN_walking",
    "MIN_biking", "MIN_PM", "MIN_PT", "nb_trips",
    "nb_walking", "nb_biking",  "nb_PM",  "nb_PT", "over",
    'intersec_res.1000',  'pdens_res')] <- ""
pred <- ini$pred # since in egt.ini the id variables are constant, the pred is already==0
vis <- ini$vis

# 1.3 actual imputation of EGT dataset
# Method is Random Forest, 5 imputations, 100 trees per imputation
mi.rf <- mice(egt, m=5, pred=pred, meth=meth, ntree=100, vis=vis)



##########################################################
# F Negative binomial regression on multiple imputation dataset
##########################################################

library(MASS)
# Fit the model for each of the 5 data sets
fit.nb <- with(mi.rf, glm.nb(mvpa.day.int ~
                          (nivetude_sim  + rvnu.1000)^2 +
                          (rvnu.1000 + pos.motorized)^2 +
                          (rvnu.1000 + dist_pt_res.1000)^2  +
                          educ_res.m + dest_res.1000
                         + age.10 + homme + emploi_sim ))

# Pool the results for the 5 data sets
pnb <- pool(fit.nb)

##########################################################
# G Plots of interaction effects
##########################################################
# 1. Use 'typical' values for variables
#   These values are used for the plots
#   where the variables are not of interest
#   e.g. mean distance to a public transport station will used
#   for the effect plot 'education*income'

# 1.1 typical values for factors:
#   proportions in all categories but the reference category
#   This reflects the use of the first level as the baseline level.
#   Effect Displays in R for Generalised Linear Models (John Fox);
#   journal of statistical software, Vol. 8, Issue 15, Jul 2003

m <- mi.rf$m # number of imputations
typical <- function(var, ref.level){
  Q <- U <- rep(NA, m)
  for (i in 1:m) {
    var1 <- complete(mi.rf, i)[,var]
    var2 <- ifelse(var1 == ref.level, 1, 0)
    Q[i] <- mean(var2)
    U[i] <- var(var2) / nrow(complete(mi.rf, i))  # (standard error of estimate)^2
  }
  a <- pool.scalar(Q, U, n = nrow(nhanes), k = 1)$qbar
  a
}

typ.etud2 <- typical('nivetude_sim', '2. bac - bacp2')
typ.etud3 <- typical('nivetude_sim', '3. bacp3 et plus')

typ.moto <- typical('pos.motorized', '1')

typ.empl2 <- typical('emploi_sim', '2. chomage')
typ.empl3 <- typical('emploi_sim', '3. retrait')
typ.empl4 <- typical('emploi_sim', '4. autre')

typ.homm <- typical('homme', '1. male')

# 1.2 'Typical' values for continuous variables: means
mean.pool <- function(var){
  Q <- U <- rep(NA, m)
  for (i in 1:m) {
    var1 <- complete(mi.rf, i)[,var]
    Q[i] <- mean(var1)
    U[i] <- var(var1) / nrow(complete(mi.rf, i))  # (standard error of estimate)^2
  }
  a <- pool.scalar(Q, U, n = nrow(nhanes), k = 1)$qbar
  a
}
typ.dist <- mean.pool('dist_pt_res.1000')
typ.edre <- mean.pool('educ_res.m')
typ.dest <- mean.pool('dest_res.1000')
typ.ag10 <- mean.pool('age.10')

# 2. Creation of a new dataset
#   This dataset will be used to construct the plot.
#   This part of the script is inspired by:
#   Atkins DC, Gallop RJ. Rethinking how family researchers model infrequent
#   outcomes: a tutorial on count regression and zero-inflated models.
#   J Fam Psychol 2007;21:726-35.
#   The variables of interest have values over their full range
#   The other variables have a 'typical' value (see above)


newdata <- expand.grid(
  intercept = 1,
  nivetude_sim = c(0,1,2),
  rvnu.1000 =  seq(from=-2.5, to=3.5, by=0.01),
  pos.motorized = typ.moto,
  dist_pt_res.1000 = typ.dist,
  educ_res.m = typ.edre,
  dest_res.1000 = typ.dest,
  emploi_sim2 = typ.empl2,
  emploi_sim3 = typ.empl3,
  emploi_sim4 = typ.empl4,
  age.10 = typ.ag10,
  homme = typ.homm
)
newdata$nivetude_sim2 <- ifelse(newdata$nivetude_sim == 1, 1, 0)
newdata$nivetude_sim3 <- ifelse(newdata$nivetude_sim == 2, 1, 0)

# 3. Prediction values for the new dataset,
#    based on the negative binomial model
pred <- function(data){

  data$rvnu.etud2 <- data$rvnu.1000*data$nivetude_sim2
  data$rvnu.etud3 <- data$rvnu.1000*data$nivetude_sim3
  data$rvnu.motor <- data$rvnu.1000*as.numeric(as.character(data$pos.motorized))
  data$rvnu.di.pt <- data$rvnu.1000*data$dist_pt_res.1000
  data$rvnu <- data$rvnu.1000*1000+3481.662 #set rvnu back to original scale for plotting
purposes
  data$nivetude_sim <- 0
  data$nivetude_sim[which(data$nivetude_sim2 == 1)] <- 1
  data$nivetude_sim[which(data$nivetude_sim3 == 1)] <- 2

  data2 <- data[, c("intercept", "nivetude_sim2",    "nivetude_sim3",
                    "rvnu.1000", "pos.motorized", "dist_pt_res.1000",
                    "educ_res.m", "dest_res.1000",
                    "emploi_sim2", "emploi_sim3", "emploi_sim4",
                    "age.10", "homme",
                    "rvnu.etud2", "rvnu.etud3", "rvnu.motor", "rvnu.di.pt")]
  l <- t(data2)
  # below: Coefficient and variance-covariance matrix are used
  #        to predict point estimates and confidence bands
  predict.data <- data.frame(matrix(c(pnb$qbar %*% l,
                             pnb$qbar %*% l - 1.96 * sqrt(diag(t(l) %*% pnb$ubar %*% l)),
                             pnb$qbar %*% l + 1.96 * sqrt(diag(t(l) %*% pnb$ubar %*% l))),
                           ncol=3, dimnames=list(NULL, c("Estimate", "LL.95", "UL.95"))))
  data[c("Estimate","LL.95","UL.95")] <- predict.data[c("Estimate","LL.95","UL.95")]
  data$Estimate <- exp(data$Estimate)
  data$LL.95 <- exp(data$LL.95)
  data$UL.95 <- exp(data$UL.95)
  data
}

plotdata1 <- pred(newdata)

# 4. Create plotting function
#   This function enables interaction plots for a continuous variable
#   and a continuous or categorical variable.
#   For the latter, 2 or 3 values can be chosen
plot.int <- function(data, var, value1, value2, value3=NA){
  plot(Estimate ~ rvnu, data=data, type="n",
       ylim=c(min(data$LL.95)-1,max(data$UL.95))+0.5,
       xlab = "Household income",
       ylab= "Minutes T-MVPA", cex.lab=1,cex.axis=0.75) #
  # plot interval slope group 1
  with(subset(data, data[,which(names(data) == var)] == value1), {
    lines(rvnu, LL.95, lty=1)
    lines(rvnu, UL.95, lty=1)
    lines(x = rvnu, y = Estimate, lty=1, lwd = 1)
  })
  # plot interval slope group 2
  with(subset(data, data[,which(names(data) == var)] == value2), {
    lines(rvnu, LL.95, lty=3)
    lines(rvnu, UL.95, lty=3)
    lines(x = rvnu, y = Estimate, lty=3, lwd = 1)
  })
  # plot slope group 3
  if(!is.na(waarde3)) {
    with(subset(data, data[,which(names(data) == var)] == value3), {
      lines(rvnu, LL.95, lty=5)
      lines(rvnu, UL.95, lty=5)
      lines(x = rvnu, y = Estimate, lty=5, lwd = 1)
    })
  }
}

# 5. Create JPEG file and apply plotting function
fig <- "C:/.../graph/"
jpeg(paste(fig, 'plot int education level - income.jpg', sep=''), width = 8.5, height = 8.5, units = "cm",  res = 500, quality = 150)
plot.int(plotdata1, 'nivetude_sim', 0, 1, 2)
legend(1000, 16.15, c("No secondary","Secondary - low tertiary","Higher tertiary"),
       lty=c(1,3,5), cex=0.45)
dev.off()
```