POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Advanced GPS data analysis for Activity-Based Model Estimation

TESI MAGISTRALE IN MOBILITY ENGINEERING – INGEGNERIA DELLA MOBILITA'

AUTHOR: Francesca Nadalini

ADVISOR: Prof. Pierluigi Coppola

CO-ADVISORS: Eng. Davide Floridi, Eng. Marco Trolese

ACADEMIC YEAR: 2024-2025

Introduction

Activity-Based approaches can be adopted to analyze travel behavior as a consequence of individuals' daily activities patterns using GPS data and to estimate advanced travel demand models. The development of Activity-Based Models (AcBMs) [1] requires access to large and detailed datasets, and their quality and richness are fundamental for the reliability of the final result. GPS data represents an advanced source that meets these requirements. Nevertheless, working with this kind of data also implies addressing several challenges, and specialized processing techniques are required.

In this context, the aim of the present study is to develop a method for analyzing a large GPS dataset collected from mobile phones, in order to generate a high-quality dataset suitable for the estimation of an AcBM.

This is part of a broader project carried out by the consultancy firm GO-Mobility, whose goal is to develop a comprehensive AcBM for Milan following the framework of ActivitySim, an open-source activity simulator developed in the United States [2].

1. Data Collection

Activity-Based Models work at an individual-specific level, allowing for a more disaggregated analysis of transport policies [3]. In fact, AcBMs rely on input data called synthetic population, which represents real-world individuals in the study area, coupled with a sequence of activities that they need to perform within a specific period of time. While the synthetic population is generally generated from statistical data such as census records, the activity diaries of the individuals derive from real-world tracking data, which can come from different sources.

Traditionally, Household Travel Surveys were used. However, they present notable limitations such as recall bias and high costs. Technological advancements have enabled the use of the big data, such as cell-tower data, GPS data and smart card records. Among these, GPS data is the most suitable for the present study, as accurately defining where and when individuals perform

activities requires high spatial and temporal precision [4].

The study input data consists of GPS records obtained from smartphone application aggregators. The main challenge of dealing with this kind of data is handling high variability in the frequency of the GPS pings, both across users and within the same user over time. The total number of pings analyzed is about 100 million, representing 924.000 individuals. For each individual, the information available varies from 1 to 30 days. The window of observation spans from the 15th of November 2024 to the 15th of December 2024, and only people having at least one ping in the city of Milan are considered.

2. Data analysis and trip purpose detection

The process of data analysis in which the activity diaries are constructed starting from the GPS pings consists of six different steps.

First, the initial raw dataset is cleaned, removing anomalous pings and unusable users. The anomalous pings are recognized by the calculation of the speed between consecutive pings belonging to the same user, and removing the ones registering a speed higher than 300 km/h. For unusable users, only the user-day combinations in which an individual records at least one GPS ping per hour for a minimum of three hours during the day are retained.

Then, the second step consists in recognizing the stop points and the trip points for each individual. After reviewing the approaches available in the literature, a methodology is proposed that eliminates the need to arbitrarily define speed and/or time thresholds. The methodology applies the clustering algorithm DBSCAN to the pings' coordinates weighted by the speed between consecutive pings within the same user. Figure 1 shows how the clustering distinguishes between stop points (grey dots) and trip points (light blue dots). Since some minor errors were encountered in this specific phase, some postprocessing is also needed:

1. The isolated stop points are turned into trip ones
2. The misclassified trip points are turned into stop ones using KNN

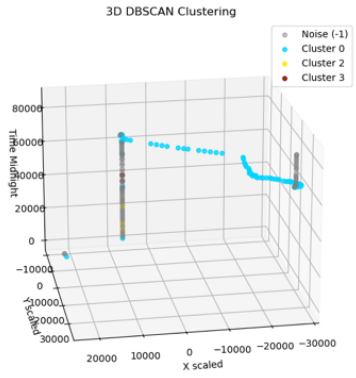

Figure 1 - DBSCAN algorithm on GPS pings

Then, the third step involves a first aggregate activities' classification: Mandatory (such as studying and working), Occasional (including all the other possible discretionary activities) and Home. The distinction is carried out applying K-Means clustering algorithm on two parameters, calculated for each activity, as to say, for each recognized stop location of each user [5]:

- $Freq_{rel,norm}$ – normalized value of the ratio between the absolute frequency of visit of the same place divided by the total number of activities performed by the user in the period of study
- $\%T_{norm}$ – normalized value of the ratio between the total time spent in a specific location and the total time the user spent performing activities in the period of study

Figure 2 shows the results of the clustering algorithm applied to the users' activities' locations. Green dots represent Home activities, yellow dots represent Mandatory activities, and purple ones represent Occasional activities.

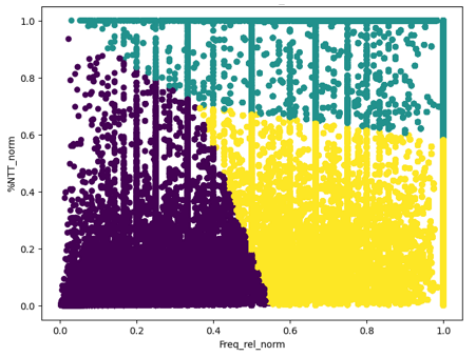

Figure 2 - K-Means on users' stop locations

The consecutive step involves the grouping of activities into tours: chains of trips that start and

end at home. The tour unit is necessary for the further AcBM estimation.

Then, a second and more detailed activities classification is performed. Land use data downloaded from OpenStreetMap (OSM) is integrated, with both the scope of giving a more detailed definition of the activities and adjust any misclassified activity, based on the combination of land use information and activity duration. Before integrating it, the OSM data is subdivided into meaningful categories: Education, Eat Out, Sports, Open Leisure, Culture and Entertainment, Shopping, Maintenance, Health Care, Services, Workplace and Residential. As regards Mandatory activities, if the geographic overlay between the activity point and the land use data corresponds to "Education", then the Mandatory activity is classified as "studying", otherwise it is classified as "working". Specific rules are applied to identify people working in education facilities.

The final step involves the decision-makers' categories identification. With all the activity diaries information, it is indeed possible to recognize some Person Types (or PTypes), following the ActivitySim structure:

- Full-time workers
- Part-time workers
- Secondary school students
- University students
- Retired people/Non-working adults

Being the socioeconomic information not available, it is not possible to define if a person is retired or simply not working, and for this reason they are grouped in the same category.

3. Aggregated statistics and data validation

In order to check if the obtained activities dataset is reliable, it is useful to analyze some aggregated statistics, as well as compare the results with similar studies.

First of all, it is important to acknowledge that the steps previously described involved a loss of data. As in every analysis, a balance had to be achieved between retaining valid information and discarding data that lacked sufficient detail for the final accuracy and reliability of the results. Figure 3 shows how the number of users changed over time, ending with 17.000 people whose data was considered reliable.

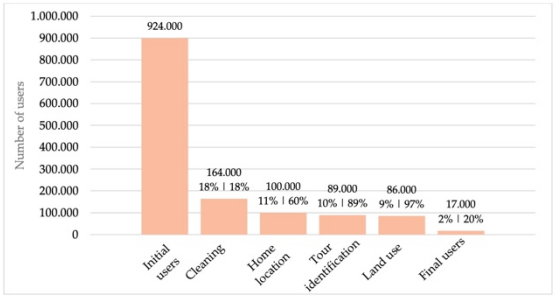

Figure 3 - Evolution of the number of users

An insightful aggregated statistics regards the distribution of the activities by starting hour, with a focus on the first classification (Mandatory, Occasional and Home) and on the difference between weekdays (Figure 4) and weekends (Figure 5). The results report satisfactory insights.

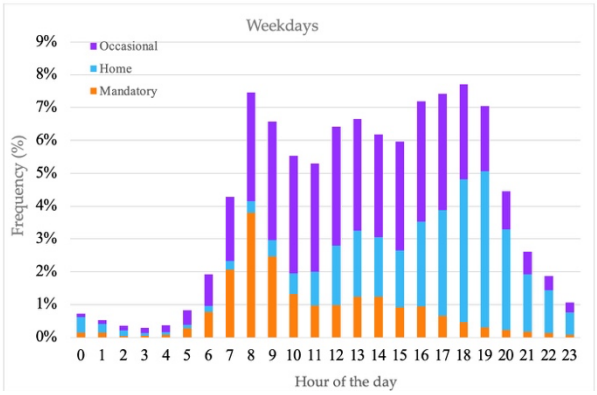

Figure 4 - Distribution of activities by starting hour during weekdays

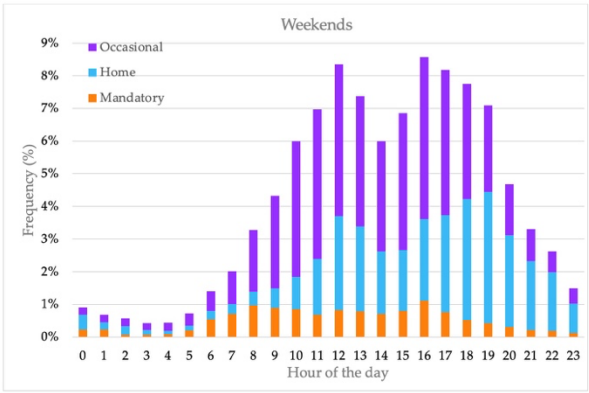

Figure 5 - Distribution of activities by starting hour during weekends

Another useful statistic concerns the average daily duration of Mandatory activities for each PType, along with their variance, both expressed in hours. Figure 6 shows these values, which are consistent with expectations.

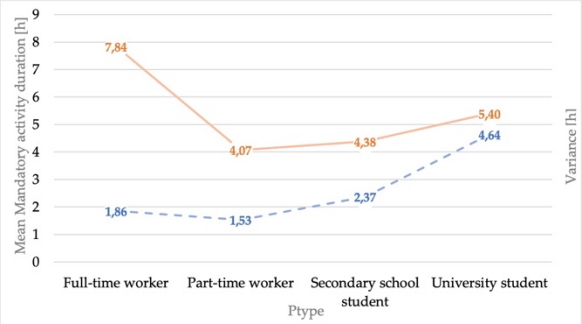

Figure 6 - Mean Mandatory activity duration and variance for each PType

From the analysis, it results that the mean number of activities per day per person, excluding the Home activity, is about 2,05, with Wednesday recording the highest value and Sunday the lowest.

Regarding instead the tour statistics, the 69% of them consists only of one stop, and the most common tour type for all the users except for non-working people is Home-Mandatory-Home.

Finally, the data is validated by comparing the results of the users recognized as Politecnico di Milano (POLIMI) students with the ones of a similar study, that was conducted analyzing the data coming from the *MotionTag* smart app [6]. Figure 7 shows the distribution of activities by their starting time for the POLIMI students of the *MotionTag* study, while Figure 8 shows the one related to the present study. The distributions are extremely similar, with the greatest difference found in the height of the peaks of Mandatory activities.

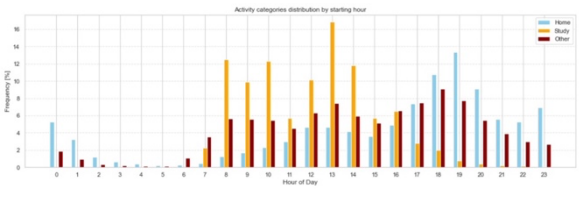

Figure 7 - Distribution of activities by starting hour for POLIMI students, present study

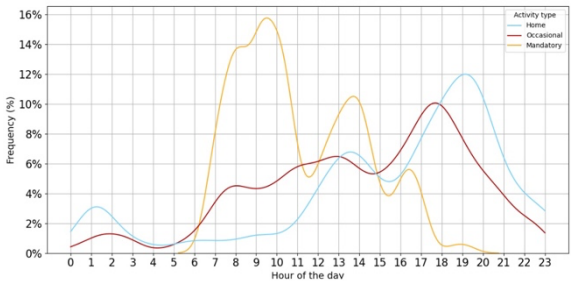

Figure 8 - Distribution of activities by starting hour for POLIMI students, MotionTag research

4. Model development

The models estimated in the present study are two of the ActivitySim sub models: an Accessibility Model for the city of Milan and the Coordinated Daily Activity Pattern (CDAP) Model. The CDAP Model is based on a typical weekday, and the choice alternatives are:

- *Mandatory (M)* – if the person engages in at least one Mandatory activity during the day
- *Non-Mandatory (NM)* – if the person engages only in discretionary activities during the day
- *Home (H)* – if the person does not go out of his/her house

The model development follows an iterative trial-and-error process, in which the model specification, calibration and validation steps are repeatedly refined until the model adequately represents observed behaviour [7].

The specification step involves the decision of the model's functional form and the attributes that will be used. The theoretical basis of the models is the Random Utility Theory, which is based on the hypothesis that every individual is a rational decision-maker who wants to maximize the utility relative to his or her choices. The utility (Equation (5.1)) is given by the sum of a systematic utility, V_j^i, which results from the attributes multiplied by their estimated coefficients (Equation (5.2)), and a random residual, ε_j^i, that represents the unobserved factors influencing individual choices.

$$U_j^i = V_j^i + \varepsilon_j^i \quad \forall j \in I^i \tag{5.1}$$

$$V_j^i(X_j^i) = \sum_k \beta_k X_{kj}^i \tag{5.2}$$

It is not possible to predict with certainty which alternative the individuals will choose, but it is possible to assess with which probability he/she will choose a specific one.

In this study , Multinomial Logit Model (MNL) and Nested Logit Model (NL) specifications have been estimated. In the MNL Model the alternatives are completely independent, while in the NL Model a correlation is introduced among groups of alternatives that users may perceive as similar. While the MNL model uses a single parameter (θ), usually set to 1, NL models include a θ_b for groups and a θ for subgroups, with θ always smaller than θ_b, reflecting the lower variance within each group.

The consecutive calibration step involves the estimation of the model's parameters, which is obtained by maximizing the so-called Likelihood function, which expresses the joint probability of observing all the choices made by the sample of users (Equation (5.3)).

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1\dots n} p^i\,[j(i)](\boldsymbol{X}^i, \boldsymbol{\beta}, \boldsymbol{\theta}) \tag{5.3}$$

There are several tools available for model calibration. One example is *Apollo*, a powerful free package for *R* [8].

The final step involves both, Informal and Formal Tests for the validation of the results. Informal tests include the critical analysis of the signs of the coefficients or their relative values for example. Formal tests, instead, involve analyzing the statistical significance of the coefficients using the T-test (Equation (5.4)), as well as evaluating the rho-squared statistics. The latter allows for comparing the estimated model with both a null model and a model that reproduces the observed shares (Equation (5.5)).

$$t = \frac{\beta_k^{ML}}{Var(\beta_k^{ML})} \tag{5.4}$$

$$\rho_{eq}^2 = 1 - \frac{\ln L(\boldsymbol{\beta}^{ML})}{\ln L(*)} \tag{5.5}$$

The key attributes used are the PType and the accessibility measures of the users' residential zones. The expressions of the systematic utilities of the choice alternatives are the following (Equations (5.6), (5.7), (5.8)):

$$V[M] = \beta_{M_w} \cdot (ftw + ptw) + \beta_{M_s} \cdot (ss + us) + \beta_{M_{ACC_{us}}} \cdot us \cdot ACC_{COLL_{TPL}} \tag{5.6}$$

$$V[NM] = \beta_{NM_{rnw}} \cdot rnw + \beta_{NM_{ACC_{RET}}} \cdot ACC_{RET_{TPL}} \tag{5.7}$$

$$V[H] = ASC_H \tag{5.8}$$

Being

ACC = the active accessibility of the users' residential zones calculated using the Gravity-Based Accessibility Measures (GraBAM) [9], as shown in Equation (5.9).

$$ACC\,(o) = \sum_i [Potential(i)^{\alpha_1} \cdot \exp\,(\alpha_2 \cdot impedance(o,i))] \tag{5.9}$$

Where *Potential(i)* refers, for example, to the number of employees in zone *i*, while the travel cost (*impedance* function) represents the travel time using different modes of transport.

The same specification is tested using first a MNL Model, then two different NL Models. Table 1 reports the estimation of the coefficients for the MNL Model along with their significance, while Table 2 reports the same model's indicators.

Table 1 - MNL Model's coefficients

Coeff.	Estimate	s.e.	t-ratio	Signif.
$\boldsymbol{\beta_{M_w}}$	2,5130	0,17154	14,650	100%
$\boldsymbol{\beta_{M_s}}$	1,5679	0,17441	8,990	100%
$\boldsymbol{\beta_{M_{ACC_{us}}}}$	0,1566	0,10114	1,548	87,9%
$\boldsymbol{\beta_{NM_{ACC_{RET}}}}$	1,2519	0,21572	5,803	100%
$\boldsymbol{\beta_{NM_{rnw}}}$	0,4091	0,02495	16,397	100%
$\boldsymbol{ASC_H}$	0,5597	0,17230	3,248	100%

Table 2 - MNL Model's indicators

Measure	Value
LL(0)	-39970,98
LL(C)	-28531,54
LL final	-28020,53
ρ_{eq}^2	0,2990
$adjusted_{\rho_{eq}^2}$	0,2988
ρ_{obs}^2	0,0179
$adjusted_{\rho_{obs}^2}$	0,0178

Apart from one coefficient, all the others present statistical significance. Looking at the indicators, instead, ρ^2_{eq} reports a good value, while ρ^2_{obs} is quite low. The attempt is to try to increase this value testing the NL Models. The NL models are tested using two configurations. In the first one, a nest groups together the NM and M alternatives, both of which involve leaving home. However, the estimated θ parameter in this case is higher than θ which contradicts the theoretical expectations of Nested Logit models. The second configuration, instead, groups together the NM and H alternatives, both of which are uncommon on a typical working day and shows more promising results. The model's performance indicators are very similar to those reported in Table 2; however, several attributes are found to be not statistically significant.

Overall, the first MNL Model and the last NL Model are the ones with the most promising results, with good values for ρ^2_{eq}, and the addition of other attributes may improve their predictive accuracy. However, at this level of choice, no further attributes could realistically be included without integrating another type of data.

Conclusion and future developments

Today AcBMs remain an active field of research, being their potential very high. This is also possible thanks to the advent of advanced data collection tools, which are able to provide high-resolution volumes of mobility data. However, sometimes this data is not collected with the purpose of estimating AcBMs, and this introduces both challenges and opportunities.

This research has addressed one of these challenging sources of data, developing a methodology to process GPS pings into a format suitable for AcBMs. The methodology has been designed to be transferable: it could be applied to other context where similar GPS-based datasets are available or even repeat it in the future for Milan. This highlights the importance of the study, which allows to save time and to reduce costs in future data preparation processes. In this sense, a logical first step for future research could be replicating the study in another city.

One key challenge was the validation of the processed dataset. Being the direct user validation impossible, aggregated statistics and POLIMI students' results comparison have been used to prove the reliability of the study. A potential future step could be further improving the validation, introducing additional external data.

As regards instead the usability of the final activities dataset, the estimation of the CDAP Model has demonstrated how it can be effectively adopted. Looking at the model indicators, it has proven to perform quite well. However, including additional data directly related to individual users, could further improve its predictive accuracy. Examples of integrable data are HTS data and census data. In both cases careful processing is required, in order to integrate the two data sources in a consistent manner.

In conclusion, this thesis has proposed a reliable methodology for analyzing GPS data to derive individuals' activity diaries and estimate an Activity-Based Model. The study underlines the relevance of such models for efficient transport planning, while recognizing that research in this field is still evolving and future developments may further strengthen the proposed approach.

References

[1] M. Ben-Akiva and J. Bowmann, "Activity Based Travel Demand Model Systems," in *Equilibrium and Advanced Transportation Modelling*, Boston, MA, Springer, 1998, pp. 27-46.

[2] E. Galli, S. Eidenbenz, S. Mniszewski, C. Teuscher and L. Cuellar, *ActivitySim: a Large-scale Agent-Based Activity Generation for Infrastructure Simulation,* San Diego, Los Alamos, National Laboratory, 2008.

[3] T. Atousa, R. Geoffrey, K. Kara and L. Hai, "Recent progress in Activity-Based Travel Demand Modeling: Rising Data and Applicability," in *Models and Technologies for Smart, Sustainable and Safe Transportation Systems*, Melbourne, Australia, and Austin, Texas, USA, 2020.

[4] S. A. Mueller, S. Paltra, J. E. R. Rehmann and K. Nagel, "Comparing GPS and cell-based mobile data to identify activity participation

during the COVID-19 pandemic," *EPJ Data Science*, 2024.

[5] L. Barbierato, *Machine Learning techniques for the analysis of people mobility,* Milano: Politecnico di Milano, 2021.

[6] J. Wolff, *Activity-Based Models estimation using smart app data,* Milano: Politecnico di Milano, 2025.

[7] E. Cascetta, Transportations systems analysis: models and applications, Springer, 2009.

[8] S. Hess and D. Palma, "Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application," *Journal of Choice Modelling*, vol. 32, 2019.

[9] P. Coppola and E. Papa, "Gravity-Based Accessibility Measures for Integrated Transport-Land Use Planning (GraBAM)," in *Accessibility Instruments for Planning Practice*, COST Office, 2012, pp. 117-124.