Dynamic Purpose Decomposition of Mobility Flows Based on Geographical Data

Etienne Thuillier[1], Laurent Moalic[2], and Alexandre Caminada[3]

1 **UTBM, OPERA, Belfort, France**
etienne.thuillier@utbm.fr
2 **Université de Haute-Alsace, LMIA EA 3993, Mulhouse, France**
laurent.moalic@uha.fr
3 **UTBM, OPERA, Belfort, France**
alexandre.caminada@utbm.fr

Abstract

Spatial and temporal decomposition of aggregated mobility flows is nowadays a commonly addressed issue, but a trip-purpose decomposition of mobility flows is a more challenging topic, which requires more sensitive analysis such as heterogeneous data fusion. In this paper, we study the relation between land use and mobility purposes. We propose a model that dynamically decomposes mobility flows into six mobility purposes. To this end, we use a national transportation database that surveyed more than 35,000 individuals and a national ground description database that identifies six distinct ground types. Based on these two types of data, we dynamically solve several overdetermined systems of linear equations from a training set and we infer the travel purposes. Our experimental results demonstrate that our model effectively predicts the purposes of mobility from the land use. Furthermore, our model shows great results compared with a reference supervised learning decomposition.

1998 ACM Subject Classification G.1.3 Linear Systems

Keywords and phrases Human mobility, Purpose decomposition, Information extraction, Linear model

Digital Object Identifier 10.4230/LIPIcs.TIME.2017.20

1 Introduction

Human mobility is a field of research that has significantly been studied during the last decades. We now understand that individuals' displacements are motivated by several factors such as jobs, occupations, social life, etc., but also by the nature of ground infrastructures. Hence, ground infrastructures are revealing items of human occupations over a territory. Moreover, humans develop tendencies to adopt regular mobility patterns, often linked to land use [9, 3]. With the apparition of pervasive devices over the last years, human mobility modeling has been significantly improved and allows us now a better understanding of such patterns. Call Detail Records (CDR) have rapidly been used as presence indicators in the literature [15, 2], but by nature CDR data represent dis-aggregated mobility flows, and often at a relatively small geographic scale (cells of the cellular network). Nowadays, many works propose interesting ways for mobility prediction, by using new technologies such as social networks [13, 1], or through heterogeneous data fusion and big data [14, 4].

From these studies emerges the idea of a link between land use and human mobility [3, 11, 17], and we understand that if the analysis of human mobility patterns leads to the characterization of land use, then, land infrastructures must be a catalyst for human

displacements. However, for privacy reasons, mobility flows are often aggregated by data providers, it is the case when dealing with mobile network data, counting loops, or any large scale mobility data. Such data can be spatially or temporarily aggregated, which represents a consequent loss of information. However, spatial and temporal decomposition of aggregated flows is a common issue, and has been largely studied in the recent years [12, 5, 7].

Purpose decomposition of aggregated mobility flows is a difficult and delicate problem. Knowing the end-purpose of any mobility flow helps local actors to better understand the dynamics of individuals traveling over their territories. Many related fields benefit from this knowledge; urbanization, transportation planning, commercial activities, etc. Many works have tried to tackle trip-purpose reconstitution in the last five years. However, the proposed methods are always dependent on the provided data nature. Floating Car Data (FCD) are GPS traces for vehicles, and by definition are not aggregated, as for CDR data. We can cite [18, 10, 6] whom infer trip activities from CDR, and in [8] the authors use FCD to determine travel purposes.

Assigning purposes to aggregated mobility flows is a heterogeneous data fusion problem, and we propose to inject knowledge into raw data to tackle this issue. In this paper, we propose to study the relation between land use (or ground) and mobility flow purposes. The main objective is to propose a method that allows us to decompose mobility flows into several sub-flows, each carrying a distinct mobility purpose. In section 2 we describe the data sets used and we explain the methodology developed to collect land use indicators. Then we propose in section 3 a reference model inferring mobility purposes from land use indicators. We provide two major improvements to this reference model and we analyze the prediction rates of the three algorithms. Finally, we propose in section 4 an analysis of these results, and section 5 concludes our work.

2 Data sets

In this paper, we propose to study the relationship between mobility flows and land use. We focus our study on the Ile-de-France province, a 12 million inhabitants region around Paris, France. We base our study on two freely available data sets. Mobility flows are gathered from a national transportation survey while land use indicators are obtained through several national databases.

2.1 Mobility flows data set

In this proposed model, we use a national transportation survey (Enquete Globale Transport, EGT). This national survey contains the declarations about displacements and commuting habits of more than 35,000 individuals. It also contains the main purpose, time, duration, origin and destination zones of each displacement. Moreover, each surveyed displacement is given a weight corresponding to the number of individuals it actually represents, based on its social and professional category, commuting habits, etc. From the EGT, we build an Origin-Destination matrix (OD matrix) that we call M_{global} and which corresponds to the daily displacements occurring within the whole Ile-de-France province.

2.1.1 Territorial division

The EGT is based on a territorial division whose base unit is a 100x100 meter mesh. By using such meshing, all of the 1,300 Ile-de-France cities are divided into regular squares. According to the EGT, these cities (and thus meshes) are themselves grouped into 118 sectors, which

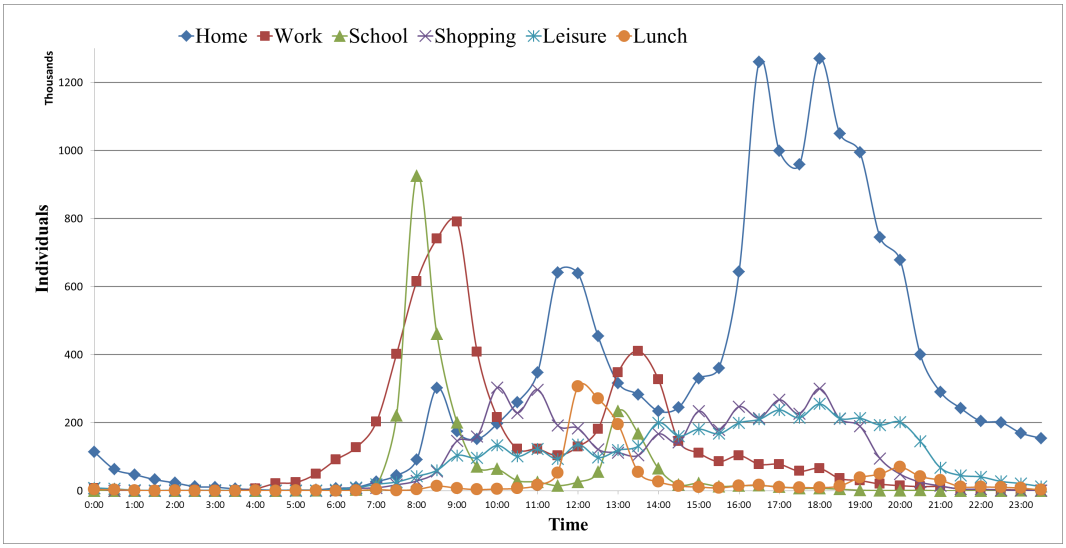

Figure 1 Number of individuals traveling with a specific purpose from the EGT.

means that a sector is composed of 11 cities on average. Origin and destination of each displacement are represented by two 100x100m meshes. This allows us to build OD matrices with any desired territorial division, from meshes to sectors. In this study we decide to use an OD matrix based on the sector division. Indeed, although the 100x100 meter meshing is interesting, it does not provide a statistically realistic information. We call the set of 118 sectors (also known as zones) Z_{global}. The origins and destinations of the M_{global} matrix belong to Z_{global}, which gives 13,924 possible OD pairs at any time.

2.1.2 Temporal division

In this study, we use temporal time slots of 30 minutes. It is a frequently used time gap which allows us to better display the mobility dynamics of the individuals. We note that in the EGT, the start and end times of each displacement are given to within a minute. To build M_{global}, we round down every time to the nearest 30 minute gap. To be more statistically correct, we also could use a Gaussian distribution model that would provide a more uniform time distribution. We refer to a timestamp of M_{global} by t.

2.1.3 Purpose division

Many works in the literature aim at inferring purposes (also called activities) of displacements from mobility data. The numbers of purposes vary greatly according to the studies. For example, in [11] the authors use nine purposes of mobility, in [18] they use five classes, and in [6] they use eight purposes. In the EGT, we have access to 38 purposes of mobility that are classified into eight main groups.

In this paper, we propose to study the six most significant distinct purposes of mobility: Home, Work, School, Shopping, Leisure and Lunch. We consider that these purposes hold the principal reasons of dynamics and movements of individuals on a territory. We notice that the proposed model can easily deal with another number of purposes. We propose to study the evolution of the EGT mobility flows, according to their purposes. Figure 1 shows the number of displacements at any time, grouped by activity.

2.2 Land use

Simultaneously, we collect land usage information from the 118 sectors of Z_{global}. In this paper, we consider that the land use of a sector is an indicator of a specific human activity done in this zone. We do not focus on the distribution of infrastructures on the ground, but we rather collect activity indicators that reflect an understanding of the usage of ground infrastructures. For example, contrarily to [17] that focuses on five land uses obtained through aerial analysis of ground infrastructures (massGIS), we propose to collect land use information from national databases. We focus on six land use indicators that are grouped into two main fields:

1. **Presence indicators**
   - Number of residents
   - Number of employees
   - Number of students (from elementary to postgraduate education)
2. **Economical activity indicators**
   - Number of megastores ($> 2500m^2$)
   - Number of supermarkets ($> 400m^2$)
   - Number of stores ($< 400m^2$)

These indicators are collected from several INSEE free-access databases. INSEE is the official French national institute for statistics and economic studies, in charge of statistics and censuses (national census, surveys, economic indicators, etc.). We collect the number of residents from the national census, which contains information about the 1,300 cities of the Ile-de-France province. The number of employees is obtained from a dedicated database[1] that nationally identifies every company, its number of employees and its location. We do not make assumptions between the different types of workers (commuting, teleworking, transporters, etc.). A version considering these special features will be checked further. We obtain the number of students from another commonly used database[2]. Finally, the number of megastores, supermarkets and stores is collected from a third database[3] that censuses every community facilities with their location.

These three economic activity indicators appear as particularly relevant since they are strong catalysts of mobility, in the sense that they do attract individuals, for identified reasons, and in different quantities. Megastores ($> 2500m^2$) mainly attract individuals for leisure, shopping or errands, in great quantities. They are strategically located over territories and are great mobility hubs. Supermarkets generate lower displacements. Individuals generally go to supermarkets to buy groceries, and more rarely for leisure. Finally, city stores ($< 400m^2$) are representative of the attractiveness and dynamics of a city center. The more city stores there are, the more individuals are present for leisure, lunch, shopping, etc. at specific times of the day. A version considering the sales volumes of these infrastructures will be checked further. As a matter of fact, some stores may be more attractive than others (services, leisure facilities, etc.) and thus may present different attraction behaviors.

We present in Table 1 some statistics about the number of indicators for the 118 zones of Z_{global}.

[1] http://www.sirene.fr/sirene/public/accueil?sirene_locale=en
[2] https://www.insee.fr/fr/statistiques/1913211
[3] https://www.insee.fr/fr/metadonnees/source/s1161

Table 1 Ground indicators details over the zones.

Indicator	min	median	max
Residents	1,177	58,732	626,676
Employees	825	23,499	365,446
Students	168	13,864	179,063
Megastores	0	1	7
Supermarkets	0	9	110
Stores	1	109	2,802

3 Model

To study the relation between land use and mobility purposes we propose to use a supervised learning model. We split our data into two parts, one part will be used for training and learning process while the second part will be used for testing and validation.

3.1 Creation of a training set

In supervised learning we have to split our database in two. The number of zones in the study being relatively small, we propose to use a 50% ratio for separating training and validation zones. With this 50% ratio we limit the risks of having too much outliers in the validation set. A version with different ratios and statistical inference models will be checked further. We propose then separate 59 EGT of the 118 EGT sectors that we put in a Z_{tr} set (tr is for training). We put the 59 other zones in a set called Z_{val} (val is for validation). All the OD flows from M_{global} with a destination zone d within Z_{tr} are added to a M_{tr} matrix, and all flows with a destination zone within Z_{val} are added to a M_{val} matrix. Additionally we create two other OD matrices called M^*_{tr} and M^*_{val}. These matrices correspond to the M_{tr} and M_{val} matrices respectively, but aggregated by destination and time slot. This means that there are no purposes information in these last two matrices. The choice of sectors is random.

3.2 Purpose Flow Decomposition algorithm (PFD)

In this paper we study the relationship between land use and displacements purposes. From one side we collect purposes decomposed flows in M_{train}, and from the other side we collect information about six land use indicators from all the zones of M_{global}. From now, we refer to a zone as a destination zone d with d in Z_{global}, and to a land use indicator at d by $ground_i(d)$. Hence, $ground_1(d)$ is the number of *Residents* and $ground_6(d)$ the number of *Stores* at zone d.

3.2.1 Normalization

In the next parts, we use the notation $n\_ < variable >$. This notation corresponds to the normalized value of a variable instance relatively to the maximum known value of this variable. For example, we use the notation $n\_ground_i(d)$ which corresponds to the normalized value of the land use indicator $ground_i$ at destination zone d relatively to the maximum known value of this ground indicator in all zones. This allows us to compare indicators between

them, without scaling effect problems. We compute this normalized value as follows:

$$n\_ground_i(d) = \frac{ground_i(d)}{\max\limits_{d \in Z_{global}}(ground_i(d))} \ . \tag{1}$$

3.2.2 Main equation

We want to mathematically write the relationship between the land uses and a purpose of mobility. For such a linear relation between the ground and mobility flows, we write for any time t, and any destination d a linear equation linking the land usage $ground_i$ and a specific purpose p:

$$\forall d, \forall t, \forall p, \ \sum_{i=1}^{n} (\alpha_i(t,p) \, . \, n\_ground_i(d)) = M_{tr}(d,t,p) \ . \tag{2}$$

where for every timestamp t, every purpose p, and every destination d,

- n is the number of ground indicators, fixed to 6
- $\alpha_i(t,p)$ is a coefficient to determine
- $n\_ground_i(d)$ is the normalized value of the ground indicator $ground_i(d)$
- $M_{tr}(d,t,p)$ is the sum of flows with purpose p from any origin zone of Z_{tr} to the destination zone d at timestamp t

3.2.3 Overdetermined system

The training matrix M_{tr} uses 59 zones, thus we can write for any couple of timestamp t of 30 minutes and purpose p, 59 equations with n unknown. As an example, we represent the specific system used for purpose *Home* and timestamp 13:30. For readability reasons we use for *n_Residents* the abbreviation *Res*, for *n_Employees*, *Emp*, etc. With same concerns, the couple $(t,p) = (13{:}30, Home)$ is removed but implicitly considered. Therefore, the values of $M_{tr}(d)$ correspond actually to M_{tr}(d,13:30,*Home*), and α_i to α_i(13:30,*Home*) coefficients. We call that system S.

$$\begin{cases} \alpha_1.Res_1 + \alpha_2.Emp_1 + \alpha_3.Stu_1 + \alpha_4.Meg_1 + \alpha_5.Sup_1 + \alpha_6.Sto_1 = M_{tr}(1) \\ \alpha_1.Res_2 + \alpha_2.Emp_2 + \alpha_3.Stu_2 + \alpha_4.Meg_2 + \alpha_5.Sup_2 + \alpha_6.Sto_2 = M_{tr}(2) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \alpha_1.Res_{59} + \alpha_2.Emp_{59} + \alpha_3.Stu_{59} + \alpha_4.Meg_{59} + \alpha_5.Sup_{59} + \alpha_6.Sto_{59} = M_{tr}(59) \end{cases}$$

As the number of ground indicators (*Res, Emp,* etc.) is equal to six, we are faced to an overdetermined linear equations system (6 unknown and 59 equations). We propose to solve these overdetermined linear systems based on a least squares approach for every timestamp t and purpose p. Since the number of equations is not large, we use a SVD decomposition as a first step for this study despite the computational cost. A version with other supervised learning models will be checked further. In the end we obtain 288 systems, where we compute n coefficients (α_1 to α_6) for any (t,p) pair.

3.2.4 Application

Now, to predict the displacements purposes, we apply to any aggregated OD flow $M^*_{val}(d,t)$, the dedicated $\alpha_i(t,p)$ coefficients. For example, to predict the *Home* sub-flow of $M^*_{val}(d,t)$, we apply the coefficients inferred from the S system trained with *Home* values at time t. For the *Leisure* sub-flow we apply the coefficients obtained with the *Leisure* values system S at

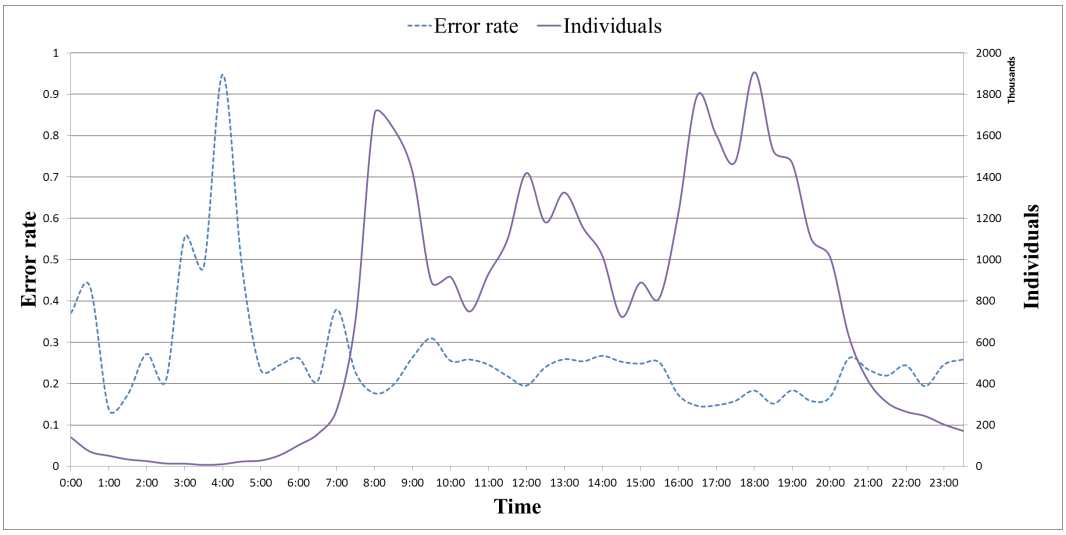

Figure 2 Error rate of the PFD algorithm.

time t, etc. It is important to note that these coefficients are not directly applied to the flow itself, but to the ground indicators $ground_i(d)$ at destination zone d. We thus obtain a theoretical displacement value $M_{theo}(d,t,p)$ for each purpose. We can then compare this theoretical value to the real displacement value $M_{val}(d,t,p)$ and estimate the error of our model. We call that model the Purpose Flow Decomposition algorithm (PFD).

Then, we write the computation of the theoretical value for purpose *Home* and timestamp 13:30 with:

$$M_{theo}(d) = \alpha_1.Res_d + \alpha_2.Emp_d + \alpha_3.Stu_d + \alpha_4.Meg_d + \alpha_5.Sup_d + \alpha_6.Sto_d \ . \quad (3)$$

For readability reasons the value $M_{theo}(d)$ corresponds in our example to the value M_{theo}(d,13:30,*Home*), and α_i to α_i(13:30,*Home*) coefficients.

3.2.5 Results

We operate Equation (3) on all aggregated flows of M^*_{val} and for all purposes *Home, Work, School, Leisure, Shopping* and *Lunch*. As a reminder, the flows of M_{val} have not been used to determine the α coefficients. We then estimate for every timestamp the number of wrongly predicted flows. For that, we compute the sum of the absolute values between theoretical $M_{theo}(d,t,p)$ and real value $M_{val}(d,t,p)$, divided by the total flow size at this timestamp. This gives us an error rate between [0, 1]. In Figure 2 we represent in dotted line the evolution of the error rate for the PFD algorithm. The solid line represents the total flow size along the day. We observe that the error rate is relatively stable over the day, except during nighttime (from 03:00 to 04:00) where the error rate jumps at 95%. During this period the number of traveling individuals is really small (around 15,000), thus the sampling is not large and individuals behavior is thus less predictable. The predicting rate of the algorithm is robust even when the number of individuals traveling increases greatly. It even reaches its optimum from 16:00 to 20:00 with almost 85% of accurate prediction whereas the number of individuals' displacements is the highest. The total daily error rate for all zones and all timestamps taken together is around 21.21%. Finally, we can state that the more individuals are traveling, the best we can predict mobility purposes.

3.3 γ-PFD, a first optimized approach

The reference PFD algorithm aims at setting down the relationship between distinct ground characteristics and different purposes of mobility. It means that for any destination d whom the land use is known, the PFD model can predict purposes of mobility with almost 78% of accuracy. We propose now to introduce in the main Equation (2) a new indicator. This new indicator considers the flow size as a determining variable. Actually, Equation (2) predicts a flow size, but do not take into account the scaling effect and flow amplitude at time t. And we see in Figure 2 that flow sizes adopt different behaviors at different times of the day. We propose then to add the total flow size $M^*_{tr}(d,t)$ as a seventh indicator in our main equation. This variable is associated to a new coefficient that we call γ. Now, by adding this new variable, one equation of the overdetermined system S becomes:

$$\forall d, \forall p, \forall t, \sum_{i=1}^{n} (\alpha_i(t,p) \,.\, n\_ground_i(d)) + (\gamma(t,p) \,.\, n\_M^*_{tr}(d,t)) = M_{tr}(d,t,p) \tag{4}$$

where

- $\gamma(t,p)$ is a coefficient to find for every timestamp t and purpose p,
- $n\_M^*_{tr}(d,t)$ is the normalized value of $M^*_{tr}(d,t)$ relatively to

  $$\max(M^*_{tr}(d,t)) \text{ for all } d \text{ in } Z_{tr} \,.$$

- and the other components are identical to the ones in equation (2).

3.3.1 Results

As for the reference PFD algorithm, we solve the overdetermined linear equations system and we compute the $\alpha_i(t,p)$ and $\gamma(t,p)$ coefficients with a least squares approach. This means that now, the equation from our example in (3) with the couple (t,p) =(13:30,*Home*) becomes:

$$M_{theo}(d) = \alpha_1.Res_d + \alpha_2.Emp_d + \alpha_3.Stu_d + \alpha_4.Meg_d + \alpha_5.Sup_d + \alpha_6.Sto_d + \gamma.n\_M^*_{tr}(d). \tag{5}$$

We operate Equation (5) on all aggregated flows of the matrix M^*_{val} and for all purposes *Home, Work, School, Leisure, Shopping* and *Lunch*. In Figure 3 we represent in dashed line the evolution of the error rate for the γ-PFD algorithm. The dotted line represents the evolution of the error rate for the reference algorithm. The total error rate over time and zones for this γ-PFD algorithm is around 17.86%. As a reminder, the total daily error rate for PFD algorithm was 21.21%. This means that the introduction of a flow amplitude coefficient increases the mean prediction accuracy of our algorithm by 3.3 points. We observe that the error rate is relatively stable over the day, except again during nighttime (from 03:00 to 04:00) where the error rate jumps at 70%. However, by introducing the flow amplitude coefficient we reduce the error during that period by almost 25 points. The γ-PFD algorithm reaches a maximum prediction rate during the period [18:00, 20:00] with nearly 90% of good prediction.

3.4 γ-PFD*, a second optimized approach

The underlying effect of solving an overdetermined system of linear equations is that the generated coefficients are adapted to give the best average solution from a training set. This means that the solver tries to give the best solutions taking into account the all

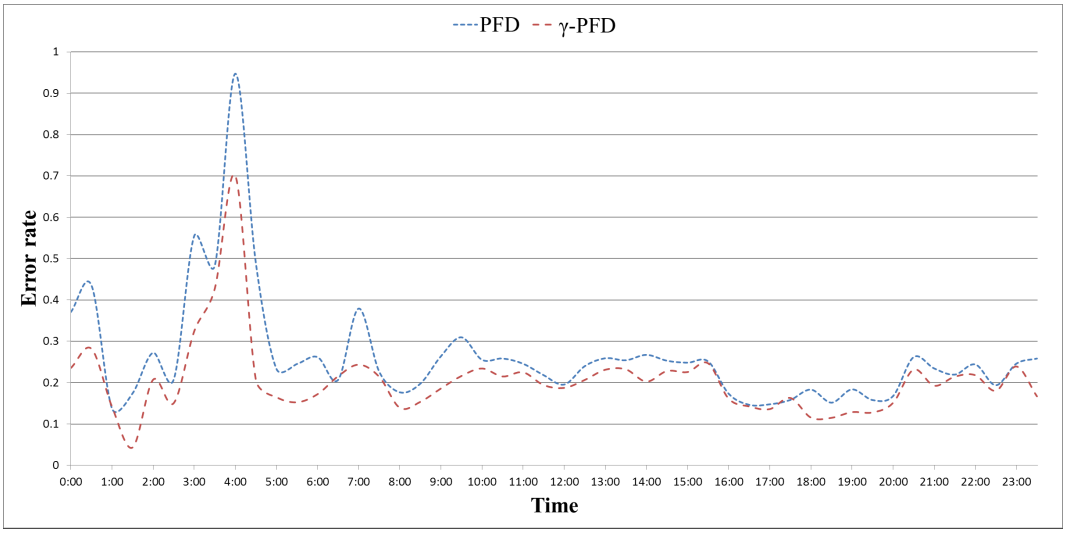

Figure 3 Error rate of the γ-PFD algorithm.

Table 2 Summary of the overdetermined systems S_1 and S_2.

System	S_1	S_2
Ground indicators	Residents, Employees, Students	Megastores, Supermarkets, Stores
Purposes	Home, Work, School	Shopping, Leisure, Lunch

different input variables. Here these variables (ground characteristics and flow size) cannot be compared directly, that is why we use normalized values. However, the solver here tries to link numbers of individuals (*Residents, Employees, Students*) and infrastructures (*Megastores, Supermarkets, Stores)* with purposes of mobility that allegedly are more attracted by specific land characteristics. As a matter of fact, individuals traveling with purpose *Work* will statistically be more attracted to a zone with more *Employees*. The same applies to the rest of the purposes. We propose here to split the system S into two twin systems S_1 and S_2, to differ primary mobility purposes and secondary mobility purposes. The primary set will address the purposes *Home, Work, School*, while the secondary set will be in charge of purposes *Shopping, Leisure, Lunch*. By doing so, the α coefficients will be more adapted to the mobility purposes inside their respective subset of learning data. We propose a summary of these two systems in table 2.

We proceed to the S_1, S_2 separation, and Equation (4) becomes:

$$\forall d, \forall p, \forall t, \sum_{i=1}^{q} (\alpha_i(t,p) \,.\, n\_ground_i(d)) + (\gamma(t,p) \,.\, n\_M^*_{tr}(d,t)) = M_{tr}(d,t,p) \tag{6}$$

where

- q is the number of ground indicators (3 for S_1 and 3 for S_2),
- with $p \in \{Home, Work, School\}$ for S_1,
- and $p \in \{Shopping, Leisure, Lunch\}$ for S_2.

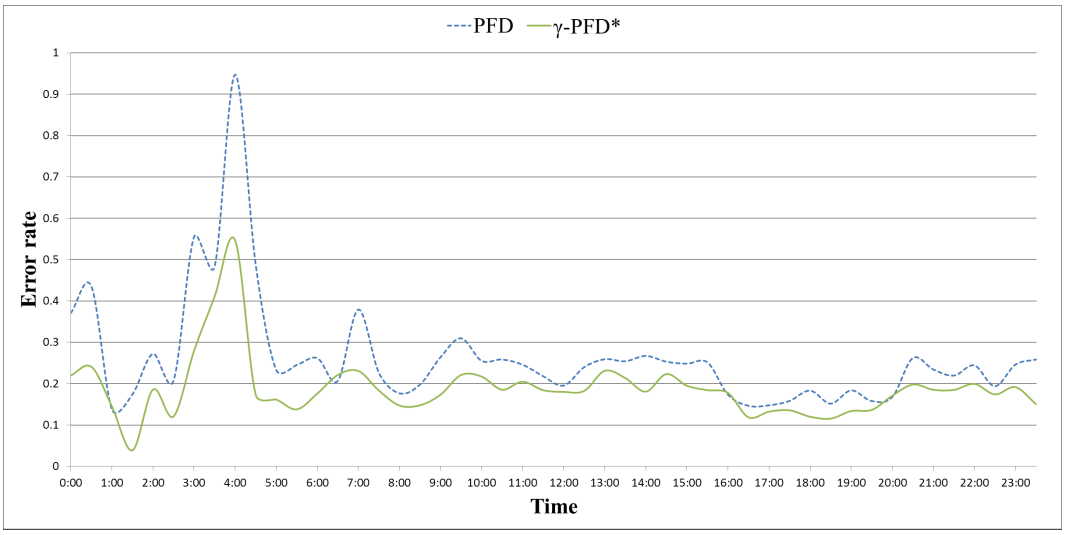

Figure 4 Error rate of the γ-PFD* algorithm.

3.4.1 Results

Now that we split our global system in two sub-systems, the equation from our example in (5) with the couple $(t, p) = (13{:}30, Home)$ is given by the equation of the system S_1 for primary mobility purposes:

$$M_{theo}(d) = \alpha_1.Res_d + \alpha_2.Emp_d + \alpha_3.Stu_d + \gamma.n\_M^*_{tr}(d) \ . \tag{7}$$

And by the system S_2 for secondary mobility purposes:

$$M_{theo}(d) = \alpha_1.Meg_d + \alpha_2.Sup_d + \alpha_3.Sto_d + \gamma.n\_M^*_{tr}(d) \ . \tag{8}$$

As for the reference algorithm PFD, we compute the $\alpha_i(d, t)$ and $\gamma(t, p)$ coefficients by solving these overdetermined systems, and we apply these coefficients on all aggregated flows of the matrix M^*_{val} and for all purposes. In other word, we apply either Equation (7) or (8) to the destination zone of M^*_{val} flows. In Figure 4 we represent in solid line the evolution of the error rate for the γ-PFD* algorithm and in dotted line the evolution of the error rate for the PFD algorithm. The total daily error rate for the γ-PFD* algorithm is around 16.84%. With this system separation we increase the prediction accuracy of our algorithm by 4.3 points.

4 Analysis

4.1 Application on the training set

The γ-PFD* algorithm gives a correct average prediction rate of 83% for the validation set M^*_{val}. We now wonder how the algorithms behave when used on their own training set M^*_{tr}. Figure 5 shows the error rates of the three algorithms when used with M^*_{tr}. We observe that all three algorithms adopt the same behavior, with an average good prediction rate of 86%. This means that the linear combinations of the ground indicators generated by the supervised learning are well adapted to the training set. However, when confronted with another set of data, introducing the flow amplitude indicator is beneficial.

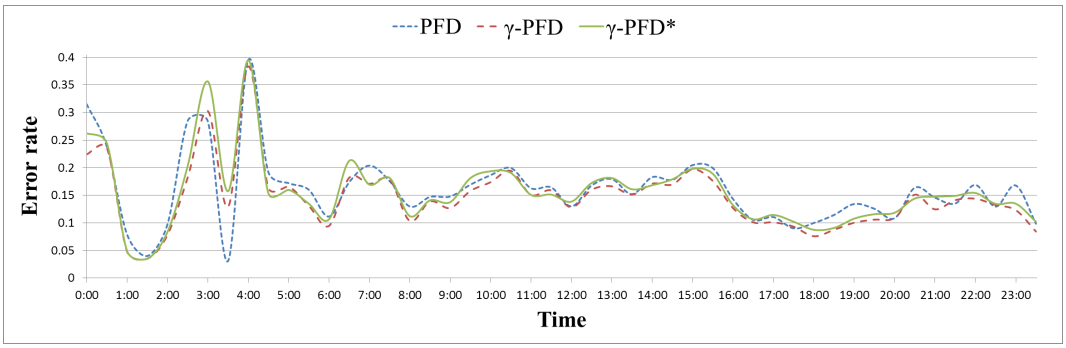

Figure 5 Error rates of the 3 algorithms when used on M^*_{tr}.

4.2 Geographical analysis

As explained before, our model gives predictions based on ground characteristics. So when the model wrongly assigns purposes to individuals flows, it means that the model has been tricked by the ground characteristics of the destination zone. We propose here to study the prevalence of some zones to give wrong results. For that we study the correlation between the daily error rate of each zone and the ground characteristics for all zones in Z_{val}. Figure 6 shows these correlations. On the abscissa we represent the ground indicator values, and on the ordinate we display the daily error rates. Each point corresponds to one of the 59 zones of Z_{val}. We observe that for all land uses, the correlation curve adopts a $\frac{1}{\log(x)}$ like pattern. Next to each graph we display the Spearman's rank correlation coefficient which is adapted to describe the correlation between two variables without linear relation [16]. All curves have a good Spearman correlation (value close to -1) and show the same trend. The more a destination zone has important land indicator value, the more the prediction rate is good. In other words, the more residents, students or supermarkets in a zone, the more the γ-PFD* model accurately predicts the purposes of mobility. We note that similar results are obtained from the training set Z_{tr}.

To validate this point, we apply the γ-PFD* algorithm on the zones having more than 40,000 inhabitants. In the *Residents* figure a vertical line shows this 40,000 limit. As a reminder, the median number of inhabitants per zone is 58,000 in our data set. With this parameter the daily error rate with the γ-PFD* algorithm is 16.01%. The gain is not important (0.8%), but it opens an interesting way for future improvements. This zone selection has been done on the other ground indicators with similar results.

5 Conclusions

Purpose decomposition of aggregated mobility flows is a delicate problem that has recently been treated from mobile network databases analysis. In this paper, we propose three different algorithms predicting purpose distributions of aggregated mobility flows, with different prediction results. The reference algorithm that we propose uses supervised learning to infer purposes of mobility from raw ground indicators. We then propose two improvements to this reference algorithm using freely available databases. We notably add a variable linked to the mobility flow size, and we propose to split the system into two sets managing distinct purposes. *Home, Work* and *School* purposes are inferred from *Residents, Employees* and *Students* indicators, while *Leisure, Shopping* and *Lunch* purposes are inferred from *Megastores, Supermarkets* and *Stores* information.

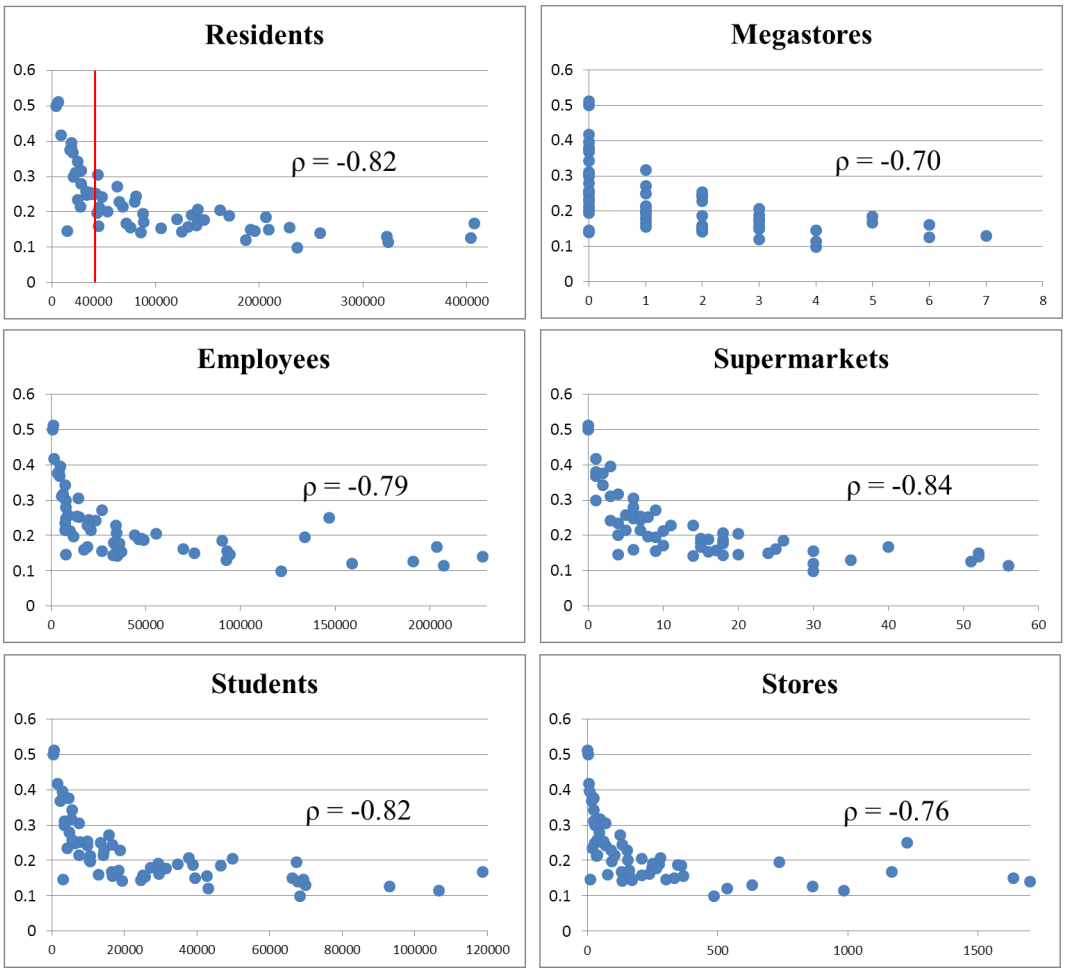

Figure 6 Correlations between ground characteristics and daily error rate.

The last improvement of the initial algorithm accurately predicts purposes of mobility in 83% of cases. And even when the number of individuals in displacement increases significantly, the prediction rate stays stable. It even reaches an optimum during the period [16:00, 18:00] with nearly 90% of success. These are promising results, as they allow a purpose decomposition of aggregated mobility flows without anything more than freely accessible sociological and geographical databases.

Furthermore, we see an interesting geographical correlation between the daily error estimation rate of the studied zones and their land use. The more infrastructures are present in a zone, the better are the prediction results. The same applies to the number of individuals moving. The more individuals are traveling, the better are the prediction results. This means that we can estimate the confidence rate of our results according to the input variables. This opens a new way for the analysis of aggregated mobility flows, especially from mobile network data.

References

1 Mariano G. Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, 5(1):30, October 2016. doi:10.1140/epjds/s13688-016-0092-2.

2 Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, number 8190 in Lecture Notes in Computer Science, pages 663–666. Springer Berlin Heidelberg, January 2013.

3 F. Calabrese, G. Di Lorenzo, and C. Ratti. Human mobility prediction based on individual and collective geographical preferences. In *2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 312–317, September 2010. doi:10.1109/ITSC.2010.5625119.

4 Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68:285–299, 2016. doi:10.1016/j.trc.2016.04.005.

5 Cathal Coffey and Alexei Pozdnoukhov. Temporal Decomposition and Semantic Enrichment of Mobility Flows. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN'13, pages 34–43, New York, NY, USA, 2013. ACM. doi:10.1145/2536689.2536806.

6 Mi Diao, Yi Zhu, Joseph Ferreira, and Carlo Ratti. Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design*, 43(5):920–940, September 2016. doi:10.1177/0265813515600896.

7 Zhanwei Du, Bo Yang, and Jiming Liu. Understanding the Spatial and Temporal Activity Patterns of Subway Mobility Flows. *arXiv:1702.02456 [cs]*, February 2017. arXiv: 1702.02456.

8 Li Gong, Xi Liu, Lun Wu, and Yu Liu. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, 43(2):103–114, March 2016. doi:10.1080/15230406.2015.1014424.

9 Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying Important Places in People's Lives from Cellular Network Data. In Kent Lyons, Jeffrey Hightower, and Elaine M. Huang, editors, *Pervasive Computing*, number 6696 in Lecture Notes in Computer Science, pages 133–151. Springer Berlin Heidelberg, January 2011.

10 S. Jiang, J. Ferreira, and M. C. Gonzalez. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, PP(99), 2017. doi:10.1109/TBDATA.2016.2631141.

11 Shan Jiang, Marta C. Gonzalez, and Joseph Ferreira. Understanding the Link between Urban Activity Destinations and Human Travel Pattern. *MIT web domain*, July 2011.

12 M. Katranji, E. Thuillier, S. Kraiem, L. Moalic, and F. H. Selem. Mobility data disaggregation: A transfer learning approach. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1672–1677, Nov 2016. doi:10.1109/ITSC.2016.7795783.

13 Anastasios Noulas. *Human urban mobility in location-based social networks : analysis, models and applications*. PhD thesis, University of Cambridge, UK, 2013.

14 Lucas M. Silveira, Jussara M. de Almeida, Humberto T. Marques-Neto, Carlos Sarraute, and Artur Ziviani. MobHet: Predicting human mobility using heterogeneous data sources. *Computer Communications*, 95:54–68, 2016. doi:10.1016/j.comcom.2016.04.013.

15 Zbigniew Smoreda, Ana-Maria Olteanu, and Thomas Couronné. Spatiotemporal data from mobile phones for personal mobility assessment. *Transport Survey Methods: Best Practice for Decision Making*, 2013.

16 C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi:10.2307/1412159.

17 Jameson L. Toole, Michael Ulm, Marta C. González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp@KDD 2012, Beijing, China, August 12, 2012*, pages 1–8. ACM Press, 2012. doi:10.1145/2346496.2346498.

18 Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, July 2015. doi:10.1007/s11116-015-9598-x.