Lucas André de Alencar

T-PROFILES: A METHOD FOR INFERRING SOCIO-DEMOGRAPHIC PROFILES FROM TRAJECTORIES

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do Grau de Mestre em Ciência da Computação.
Orientadora: Prof. Dr. Vania Bogorny

Florianópolis

2015

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Alencar, Lucas André de
T-Profiles: a Method for Inferring Socio-Demographic Profiles from Trajectories / Lucas André de Alencar ; orientadora, Vania Bogorny - Florianópolis, SC, 2015.
64 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação.

Inclui referências

1. Ciência da Computação. 2. Perfil Sócio-demográfico. 3. Trajetórias de objetos móveis. I. Bogorny, Vania. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

I would like to dedicate this work for my brother, that has a great future ahead of him.

ACKNOWLEDGEMENT

I would like to thank my grandmother Maria Neusa Duarte de Alencar, that even being far from me, always supported and motivated me to persist and achieve my goals. To my advisor Prof. Dr. Vania Bogorny that guided me through the uncertain path to the master's degree and that mobilized a huge amount of effort for this work to be completed. Also to Prof. Dr. Luis Otavio Alvares that has always been present and participating in the discussions, where I am certain that has influenced a lot in the obtained results and the final proposal.

I must thank CAPES institution that supported the development of the work through its scholarship. To PPGCC (Programa de Pós-Graduação em Ciência da Computação) and UFSC (Universidade Federal de Santa Catarina) for the opportunity to be part of the institution which contributed a lot for my academic growth. To the Marie Curie Project and the European Union which made possible for me to have an amazing experience at CNR (National Research Council) in Pisa/Italy through the SEEK project, where I met many incredible researchers which helped me in this work. I would like to thank to two major researchers, Chiara Renso (ISTI-CNR) and Alessandra Raffaeta (DAIS - Università Ca'Foscari Venezia), participants in the SEEK project, which have contributed to this work and have helped me a lot during my staying in Pisa.

Finally, I would like to thank all my friends for being there in the moments of fun, giving me strength to continue and conclude this long journey. Just to cite some names, to all my friends from my hometown that have moved to Florianópolis with me: João Pozzobon, Ricardo Nabo, Diogo Busanello and Anderson Salles. To everyone at LISA lab: Artur Aquino, Fernando de Lucca Siqueira, André Salvaro Furtado, Areli Andréia Santos, Carlos Andres Ferrero, Juarez Angelo Piazza Sacenti, Ramon Hugo de Souza, Cleto May, Geomar Schreiner and Douglas Klein.

- Do you know the only thing happier than a three-legged dog? A four-legged dog...

(Dr. Bigelow [Louis C.K.], 2014)

RESUMO

Ter o conhecimento sobre o perfil dos habitantes de uma cidade ou país tem grande valor para administrações públicas e empresas. Conhecer o perfil de uma população pode auxiliar o trabalho de planejadores urbanos, administradores de transporte público, serviços governamentais ou empresas de diferentes maneiras como, por exemplo, decidir onde é interessante instalar uma nova loja ou personalizar anúncios para um determinado público. A forma mais comum utilizada na análise de informações demográficas de uma população é através da segmentação da mesma em perfis sócio-demográficos, como idade, ocupação, estado civil ou renda mensal. Atualmente, para que essas informações sejam descobertas e analisadas, os dados são coletados através de entrevistas realizadas de casa em casa, periodicamente, em diversos países. No entanto, este tipo de abordagem possui algumas desvantagens: 1) os dados não são atualizados e precisos, pois são coletados em um intervalo de 5 - 10 anos; 2) a coleta é muito custosa e cobre apenas uma parcela da população por um curto período de tempo, apesar de ser estatisticamente significante; 3) não caracteriza as atividades completas do indivíduo, apenas o período de 1 dia de atividades, fornecidas através da entrevista realizada. Atualmente, é possível inferir muito conhecimento a partir do comportamento das pessoas analisando seu movimento do dia-a-dia, uma vez que grandes quantidades de dados de movimento estão disponíveis como: dados de telefone celular, redes sociais, dados de GPS, etc. Nesta dissertação, é proposto um método para a extração de perfis sócio-demográficos a partir de trajetórias de objetos móveis, e apresenta as seguintes contribuições: (i) proposta de um modelo de perfil geral para representar o perfil sócio-demográfico de pessoas, como trabalhador, estudante, desempregado, etc; (ii) proposta de um modelo para representar o histórico de movimentação diária dos indivíduos; (iii) proposta de funções de similaridade para fazer o casamento entre histórico e modelo de perfil e; (iv) um algoritmo chamado T-Profiles que realiza a comparação entre modelo de perfil e modelo de histórico, com o intuito de inferir o perfil sócio-demográfico de um objeto móvel a partir de sua trajetória. O algoritmo T-Profiles é validado utilizando dados reais de trajetórias, obtendo em torno de 90% de precisão.

Palavras-chave: Perfil Sócio-demográfico. Trajetórias de objetos móveis.

ABSTRACT

The knowledge about people living in a city or country has great value for the public administration as well as for enterprises. To know the population profile may help the job of smart city planners, public transportation administrators, government services or companies in many different ways, such as to decide if and where to install a new store or to personalize an advertisement, for example. The usual approach for population demographic analysis is to segment the population in socio-demographic profiles, such as age, occupation, marital status or income. Most attempts to discover and measure the population profiles is through human surveys, and the most well-known example is the socio-demographic census with diary activities, done periodically in many countries. However, the main drawbacks of the census data is that they: 1) are not up to date since they are usually collected every 5 - 10 years; 2) are expensive to collect, and cover only a small - although statistically significant - part of the population for a short period of time; 3) do not collect the actual movement of the individuals, but only the activity performed during one day and which is mentioned by the user during the interview. We believe that nowadays we can infer much knowledge and the real behavior about people from their every day movement. In this thesis we propose a method to extract socio-demographic profiles from trajectories of moving objects, and make the following contributions: (i) we propose a general profile model to represent socio-demographic profiles of people such as worker, student, unemployed, etc; (ii) we propose a moving object history model to represent the daily movement of the object, and (iii) we propose similarity functions and an algorithm called T-Profiles for matching the profile model and the history model in order to infer the socio-demographic profile of a moving object from his/her trajectories. We validate T-Profiles with real trajectory data obtaining about 90% of precision.

Keywords: Socio-demographic profiles. Moving object trajectories.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

T	Raw trajectory	36
tid	Trajectory ID	36
p_i	Trajectory point	36
x_i	Trajectory point coordinate x	36
y_i	Trajectory point coordinate y	36
t_i	Trajectory point time	36
POI	Point of Interest	36
$_{POIType}$	Type of Point of Interest	36
$stop$	Trajectory stop	36
$\mathcal{P}$	Set of Profile names	37
p	Profile name	37
r	Profile rule tuple	37
$\mathcal{R}^p$	Profile model for profile name p	39
h	Object history	39
oid	Object ID	39
$\mathcal{M}_h$	Moving object history model for object history h	39
m	Moving object history model tuple	39
sim_{pos}	Similarity function for positive rules	41
sim_f	Similarity function for frequency	41
sim_d	Similarity function for duration	41
sim_{neg}	Similarity function for negative rules	43

SUMMARY

1 INTRODUCTION AND MOTIVATION

The knowledge about people living in a city or country has great value for the public administration as well as for enterprises. To know the population profile may help the job of smart city planners, public transportation administrators, government services or companies in many different ways, such as to decide if and where to install a new store, to personalize an advertisement, where to add new bus routes, etc.

A profile is a set of features that represent a group of people with similar characteristics. The usual approach of population demographic analysis is to segment the population in socio-demographic profiles, such as age, occupation, marital status or income. Most attempts to discover and measure the population profiles is through human surveys, and the most well-known example is the socio-demographic census with diary activities, which is done periodically in many countries. However, the main drawbacks of the census data is that they: 1) are not up to date since they are usually collected every 5 - 10 years; 2) are expensive to collect, and cover only a small - although statistically significant - part of the population for a short period of time; 3) do not collect the actual movement of the individuals, but only the activity performed during one day and which is mentioned by the user during the interview.

We believe that nowadays we can infer much knowledge and the real behavior about people from their every day movement. We are entering the era of big data, where the real movement behavior of a society can be extracted from its individuals everyday movement. We all leave electronic traces of our behavior during our life through web logs, credit card transactions, GPS devices, WiFi and GSM networks. Although we may not realize, we are constantly being tracked, generating a new type of data, called *moving object trajectory*. A moving object trajectory is a set of points that describe the movement of an object in space and time. A raw trajectory is represented as a set $\langle tid, p_0, p_1, ..., p_n\rangle$ where *tid* is the trajectory identifier and $p_i = (x_i, y_i, t_i)$ is a point, where x_i and y_i are the geographical coordinates that correspond to the place where the object was located during the instant of time t_i, given $i = 0, ..., n$ and $t_0 < t_1 < ... < t_n$. From these data it is possible to discover the behavior of people as well as their daily routine (BRAZ; BOGORNY, 2012).

In daily life, in general, we all follow a *routine*, going more or less to the same types of places everyday (e.g. work, gym, supermarket, restaurant, etc). The routine of a person during one week, one month or one year represents the *general pattern of movement* of this person. For instance, a typical routine of a *worker* is to go, in general, four or five times a week to work, while a *student* goes to school/university four or five times a week. On the contrary, an *Unemployed* may have a different routine, as not going to work everyday. The routines followed by a similar group of people as the students, workers, or unemployed we call *profiles*.

With the increasing number of GPS trajectory datasets and the definition of semantic trajectories in the domain of mobility data, introduced by (SPACCAPIETRA et al., 2008) and extended in (BOGORNY et al., 2014), it is possible to infer the places visited by an object, the duration of the visit, and the frequency of the visits. A semantic trajectory in (SPACCAPIETRA et al., 2008) is defined as a set of stops and moves, where the stops are the places that an object has visited. Based on these visits, it is possible to obtain the routine of an object. An example of semantic trajectory is shown in Figure 1, where the moving object visits four different places (home, university, shopping mall and bar), so having four stops in one day.

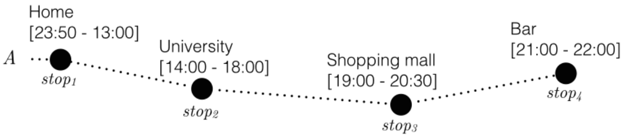

Figure 1 – Example of semantic trajectory A.

The discovery of profiles from movement data, more specifically from semantic trajectories, can reveal the real state and social behavior of a population. It can show *the real* activities and habits of a population in a much larger and realistic scale than any survey or questionnaire.

1.1 PROBLEM STATEMENT

In the literature of moving object trajectories there are several works related to the extraction of general patterns from trajectories and the summarization of the movement of objects. No works have tried to look deeper in the data to infer more knowledge about the moving object itself. Only a few works address user profiles, but from a different perspective and for different mobility data. For GPS trajectories, which is the focus of this work, Trasarti in (TRASARTI et al., 2011) defines as object profile the representative trajectory of a set of similar trips, for car pooling. No information about the trajectory is used to discover who is the moving object and which are his/her activities. The work of (XIAO et al., 2010) defines as profiles the users that visit similar places at similar time, in order to discover users with similar habits, but there is no inference about the meaning of such habits. The work of (ZHENG; ZHENG; YANG, 2009) is the only work that proposes to infer socio-demographic profiles, but for social network data integrated to GPS trajectories, not only from pure trajectory data. In summary, in these works a profile is considered as a set of features which characterize a type of user or a group of users, but not for socio-demographic inference.

The socio-demographic profile is an interesting information that may be used in different applications in order to understand the population behavior of specific places. For instance, real state companies, using the socio-demographic profiles of an area of a city or a neighborhood, can indicate the best regions for a client to live or to settle a business based on the profile of the people that reside there. Marketing companies that are working with a product focused on a specific group of people can use the socio-demographic profiles to formulate their marketing strategy. Public transportation administrators can use the profile information and mobility pattern to create new bus routes or metro stations that fit the population needs.

In this work we propose a different perspective from the previous works. We assume that an object history is given, as shown in Table 1, and a description of a mobility behavior for specific socio-demographic categories of users is available and can be represented as “rules”. These rules can be defined by domain experts who describe which is a typical behavior of a specific profile category (workers, students, unemployed) in a certain application. Another possibility is to run data mining methods on census data or on GPS trajectories to identify groups of users with similar behavior, and label them with the socio-demographic category like “workers” or “students” (JIANG; FERREIRA; GONZÁLEZ, 2012). Thus, given domain knowledge about how to describe a socio-demographic profile, we propose a profile model based on the rules that a moving object should fulfill to belong to a specific profile category. This model allows the user to specify, in a simple way, the types of profiles that are interesting for an application. How to match GPS trajectories and the profile model is the second focus of this thesis, which proposes a moving object history model and a set of similarity functions that are capable to take into account the blurred aspect of such profiles in two ways: (1) the temporal match is defined considering the overlapping portion between a profile model and the trajectories behavior; (2) the matching function assigns to the match a similarity degree between the profile model and the trajectory behavior. In other words, a trajectory may be matched to several profiles with different similarities, thus characterizing multiple profiles.

Table 1 – Example of object history.

Object ID	Type of place	Start time	End time
1	Workplace	02-03-2015 07:17	02-03-2015 11:35
1	Restaurant	02-03-2015 11:58	02-03-2015 12:49
1	Workplace	02-03-2015 13:13	02-03-2015 17:32
1	Supermarket	02-03-2015 18:03	02-03-2015 19:11
1	Home	02-03-2015 19:41	02-03-2015 07:05
1	Workplace	03-03-2015 07:22	03-03-2015 11:39
1	Restaurant	03-03-2015 12:04	03-03-2015 12:58
1	...	...	...
1	Workplace	05-03-2015 12:55	05-03-2015 17:35
1	Supermarket	05-03-2015 18:12	05-03-2015 18:46
1	...	...	...
1	Workplace	03-04-2015 07:28	03-04-2015 12:07
1	...	...	...

To the best of our knowledge, there is no previous work that infers the same type of profiles from GPS trajectory data as we propose in this work. In summary, we make the following contributions with respect to previous works:

1. We define a socio-demographic profile model as a set of rules, which describe the behavior that we want to discover;

2. We define a moving object history model for representing the summary of movement behavior of individuals from the trajectories;

3. We propose the algorithm T-Profiles to extract socio-demographic profiles from trajectory data, that is able to infer multiple profiles of a single object by matching the profile model with the history model;

4. We experiment the proposed approach with GPS trajectory data, but the method is generic in order to deal with any type of mo-

bility data which allows the extraction of the type of place that an object has visited and the time of the visit.

1.2 OBJECTIVE

The main goal of this work is to infer the socio-demographic profile from moving object trajectory data. To achieve this result, the following specific goals must be fulfilled:

- Formally define a socio-demographic profile model;
- Formally define a moving object history model;
- Define similarity measures to match the socio-demographic profile models and the moving object history models;
- Define profile models for specific socio-demographic profiles to be used in the experiments;
- Propose and develop an algorithm that identifies the socio-demo graphic profiles from moving object trajectories using the socio-demographic profile model, the moving object history model and the similarity measures.

1.3 METHODOLOGY AND STRUCTURE

The following tasks will be performed to achieve the objectives of this thesis:

1. Review the state of the art on socio-demographic research over different types of data, such as: phone call records, web logs, social networks and GPS trajectory data;
2. Formally define the socio-demographic profile model based on places visited by a person that may reveal his/her profile;
3. Formally define the moving object history model, summarizing the characteristics presented in the object movement history such as frequency, day and week periods and duration of visits to a certain type of place;
4. Define similarity measures to verify the similarity between a moving object history model and a socio-demographic profile model;

5. Develop an algorithm to identify the socio-demographic profiles from moving object trajectories, using the data structures profile model and moving object history model defined in the previous tasks;

6. Define a set of socio-demographic profiles with their respective set of rules that describe the profile behavior which will be used in the experiments;

7. Perform experiments using activity census data and the profile models defined earlier. This dataset contains the socio-demographic information and the activity diary for each object, containing the activity performed and its period of time. Due to the lack of days in the history, where each object has only one day of diary, we group objects that have the same socio-demographic profile into 14 days long histories, respecting the day of week that the diaries occurred. The algorithm is applied to these histories returning their socio-demographic profiles with their respective similarities;

8. Preprocess different moving object datasets to obtain the necessary semantic trajectories. Using the method SMoT (ALVARES et al., 2007), we compute the stops of the trajectories, match them with types of places and build the object histories;

9. Perform experiments using real GPS trajectory data with the previously defined profile models, in order to understand how the method works with a dataset of real trajectories. The method is executed over the object histories previously computed using the profile models defined earlier;

10. Compare the results of the proposed algorithm with the classification algorithms RIPPER (COHEN, 1995) and the C4.5 (QUINLAN, 1993), with the intent to analyze the precision of the proposed method in relation to well-known classification algorithms.

The rest of this thesis is organized as follows: Chapter 2 presents the state of the art on mobility profiles and socio-demographic profiles; Chapter 3 introduces the basic and new concepts for this work; Chapter 3.4 presents the algorithm T-Profiles for extracting socio-demographic profiles from trajectories; Chapter 4 presents the experimental evaluation of the method with census data, that are used as ground truth, and real GPS trajectory data; and finally Chapter 5 presents the conclusion and future works.

2 THE STATE OF THE ART

This chapter briefly describes the state of the art related to the theme of this thesis, pointing out the characteristics of the works and the main differences with our proposal.

There are many works in the literature for finding behavior patterns in *groups* of trajectories. Some of them are (CAO; MAMOULIS; CHEUNG, 2007; GIANNOTTI et al., 2007; LEE et al., 2008). (CAO; MAMOULIS; CHEUNG, 2007) propose a method for extracting periodic patterns from moving objects. A periodic pattern is a sequence of regions of interest regularly intersected by a moving object during a certain period of time. However, this work does not make use of semantic trajectories, and its focus is on extracting the periodic patterns, and not the socio-demographic information of the objects. (GIANNOTTI et al., 2007) proposes a method for extracting regions frequently visited, considering the order of the visits and the transition times. When a minimum number of trajectories have visited similar places and have similar transition times, the method considers these trajectories as a trajectory pattern. Another work related to patterns in *groups* of trajectories is (LEE et al., 2008), which proposes a method to classify trajectories according to the places they visit. If trajectories have similar destination, the method considers them as part of the same group. A summary of the most studied behavior patterns is presented in (PARENT et al., 2013), as well as some basic concepts about semantic trajectories.

The previous works are well known in the area of pattern mining in trajectories and in some sense are related to our proposal, where they identify common spatio-temporal behaviors present in the trajectories. But none of them use semantic trajectories or are interested in the socio-demographic profiles of the moving objects as our proposal.

The inference of user profiles from GPS trajectory data, which is the focus of this work, is very recent, and the majority of works have different meanings for profiles. The related works can be grouped in two main categories: works that analyze the mobility profile of individuals, and works related to socio-demographic profiles from different data sources.

The works that analyze the mobility profile, focus on the representation of the movement. For instance, Trasarti et al. in (TRASARTI et al., 2011) defines profile as a set of representative sub-trajectories that are performed regularly by an object; Hung et al. in (HUNG; CHANG; PENG, 2009) and Chen et al. in (CHEN; PANG; XUE, 2014) define profile as a sequence of regions that are frequently visited by an object; Bayir et al. in (BAYIR; DEMIRBAS; EAGLE, 2010) proposes a profile as a collection of frequent sequences of cell towers that the user has visited; and Furletti et al. in (FURLETTI et al., 2013) defines profile based on the phone call habits of the user given an area to be monitored, as for instance, a city or a neighborhood.

For the works related to socio-demographic profiles, which is the information of the identity of people, such as age, gender or occupation, have similar profile definitions. A socio-demographic profile is defined as a collection of activities, regardless the data source, that indicates the identity of a person or a group of people. Mackinnon et al. in (MACKINNON; WARREN, 2007) and Bi et al. in (BI et al., 2013) use the activity of a user in a social network and a search engine to discover their demographic information; González et al. in (JIANG; FERREIRA; GONZÁLEZ, 2012) performs the analysis of activity census data, which contains the activities performed by the respondents during one day, to extract the socio-demographic profiles; Zheng et al. in (ZHENG; ZHENG; YANG, 2009) proposes to match social networks and GPS trajectories to infer the socio-demographic profiles; and Baglioni et al. in (BAGLIONI et al., 2009) models semantic trajectories into ontologies to extract information about tourists and commuters, based on the frequently visited places.

In Sections 2.1 and 2.2 we present more details about these works, exploring their characteristics and differences with our proposal.

2.1 MOBILITY PROFILES

The works presented in this section analyze and identify *mobility profiles*, where a profile is a repetitive/frequent movement. These works do not extract socio-demographic information, but we describe them because they use the term profiles.

Trasarti et al. in (TRASARTI et al., 2011) proposes a framework that extracts the patterns of movement of an individual, called mobility profile, based on raw trajectories. By raw trajectories we mean the set of space-time points. The work defines as user mobility profile the set of representative trips performed by the object in his/her historical movement, where a profile is a spatio-temporal trajectory which is frequent in the object's movement history. The goal is to compute these patterns of movement for a car pooling system that recommends rides, considering the user mobility profile. The framework consists of 2 phases, the first one is responsible for building the user mobility profile, which is related to a single user. The second phase matches the mobility profiles to find users with similar routine trips, in order to recommend them in the car pooling system.

Hung in (HUNG; CHANG; PENG, 2009) proposes a framework to discover communities of users with similar routines using GPS traces. The work defines as object profile the sequence of regions frequently visited by the object, and objects with similar visits are clustered to infer communities of people. The object profile is represented as a *probability suffix tree* (PST) that models in a tree structure the sequences of regions visited by the object. Each node in the PST represents a region visited by the user and each node contains a table with the probabilities for the next region to be visited. The proposed framework contains 3 steps: first, it constructs the objects profiles as PST structures; second, it proposes a distance measure to compute the similarity between the user profiles; and third, using the distance measure proposed and a clustering algorithm, it identifies the communities of users that have similar moving behaviors. Both (TRASARTI et al., 2011) and (HUNG; CHANG; PENG, 2009) use as input a set of raw GPS trajectories where the object history is a set of space-time points, while we focus on semantic trajectories where the object history is a set of stops with the places the object has visited.

Works as (YING et al., 2010) and (XIAO et al., 2010) consider semantic trajectories, which is the focus of our proposal, but they do not infer user mobility profiles. Instead they propose similarity measures for semantic trajectories. Ying et al. in (YING et al., 2010) proposes a similarity measure that estimates the similarity between semantic trajectories, where the similarity is computed based on the matches of the sequences of categories of places visited between the trajectories. Improving the approach in (YING et al., 2010), Xiao et al. in (XIAO et al., 2010) proposes a similarity measure that considers not only the sequences of places but also the travel time to the place and the duration of the visit.

Similar to previous works, Chen et al. in (CHEN; PANG; XUE, 2014) proposes an approach to construct mobility profiles from trajectories and to compute the similarity between them. The work defines as trajectory pattern the mobility behavior of an object, computing the regions of interest (dense regions) of a user as stay points. A mobility profile is then defined as the frequently visited regions of interest. Afterwards, an improved similarity measure based on the one proposed in (YING et al., 2010) is used to match objects with similar mobility profiles. In the analysis, (CHEN; PANG; XUE, 2014) proposes to distinguish the mobility profiles through temporal semantics, which are specific periods of time. For example, weekday/weekend or daytime/night. This way, the proposal is able to identify behaviors that may be hidden when considering the whole user movement history. Even though the work proposes this separation through temporal semantics, the main focus is to group users with similar movement patterns and not to discover socio-demographic profiles.

In GSM data management, a user profile is defined as his/her mobility pattern based on phone calls. An example is the work (BAYIR; DEMIRBAS; EAGLE, 2010), which extracts the mobility pattern of GSM users, considering their approximate location based on the GSM cell towers. The work defines as mobility profile a collection of the most frequent sequences of cell towers visited by a user. Another work is proposed by (FURLETTI et al., 2013), where the authors analyze GSM calls taking into account both position and time of each call to infer the status of a user as a resident, a commuter, or a visitor of a given monitored area. In this work, some rules are defined to describe the behavior of the residential users. For instance, it is expected that a resident performs calls during weekdays and weekends and a commuter only during weekdays, because a commuter only works in the monitored area during weekdays, while a resident works and lives there, making calls during the whole week. The idea is to mine the GSM calls to identify groups of users with similar call habits, and assign the residential label to these groups. (DASH et al., 2014) proposes to infer both home and workplace from GSM data, but not for profile inference.

2.2 SOCIO-DEMOGRAPHIC PROFILES

Most of the works presented in this section were developed for social media data, where the socio-demographic information is available in the data. These proposals extract socio-demographic information from different data sources to obtain information about the identity of the user such as age, gender, occupation, etc. We detail these works because they have similar goals to our approach, but as stated before, most information is present in the data, while in GPS trajectories there are only space-time points. In web log data and social networks, the inference of profiles has been an active area of research. For instance, (BI et al., 2013) proposes a model to infer demographic traits as age and gender of users based on their use of search engines and social networks, using Facebook likes and user demographic information. (MACKINNON; WARREN, 2007) explores the relationship between users with the intent to make a prediction of users age and country of residence, based on the information given by their friends on a social network.

By mining *activity census data*, González et al. in (JIANG; FERREIRA; GONZÁLEZ, 2012) performs an analysis in order to understand how humans allocate time to different activities during the day as part of a spatio-temporal socio-economic system. Through the analysis, the work intended to address the following issues: (1) what is the daily activity structure of individuals in a city, (2) what is the variation of the activities over time, and (3) how individuals can be grouped based on their activities and what socio-demographic information can be found in these groups. They perform a number of analysis on census data using clustering techniques, to identify groups of users with similar activities, and manually label the groups as students, workers, early-bird workers, stay-at-home, etc; based on the characteristics of the clusters. By statistically analyzing the individuals information available in the census, they were able to identify the socio-demographic signatures of each cluster found. For instance, how much is the proportion of females in the stay-at-home cluster, or what is the average age of the individuals in the workers cluster. In (JIANG; FERREIRA; GONZÁLEZ, 2012), González et al. proves that it is possible to group people based on their similar activities, which is one of the assumptions in our proposal. The analysis is performed in a sample of a population, which is a disadvantage compared to our work, that considers the real movement of the individuals. Besides, it does not propose a method to infer socio-demographic profiles, as we propose in this work.

Baglioni et al. in (BAGLIONI et al., 2009) proposes a model to identify frequent patterns in trajectories and classify the trajectories in profiles. However, mainly tourists and commuters are identified by this approach, since the work only considers as a profile a set of types of places that an object visits, without considering temporal information. For instance, an object that visits touristic places and hotels is considered a tourist. Different from our proposal, to assign a profile, (BAGLIONI et al., 2009) considers only the places visited by the trajectory, while we analyze the places including the temporal information such as the duration or the period of the day the object visited the place. The focus is to provide a model for conceptual representation and deductive reasoning of trajectory patterns, not to find the socio-demographic profiles.

(ZHENG; ZHENG; YANG, 2009) proposes a method that infers the user activities (e.g. studying, working, etc) and profiles (e.g. student, worker, etc) by matching GPS trajectory data, social network profiles (such as occupation, gender or age), and diary of activities that contains the activity and the period of time that the activity is performed. As the diary of activities and the social network profiles are manually annotated by the user, the work assumes that the information is incomplete, i.e., some fields in the social network profiles are not filled up or some activities from the diary were omitted. So, the proposed method infers the missing activities using the user current location and the information in the social network profile. For instance, if the user is at a university and has in his/her social network profile the occupation as student, the method infers that the activity is studying. Afterwards, the method infers the missing information in the social network profiles based on the similarity between the users diary of activities, i.e. if two users have similar activities, they probably have similar profiles. The information about the user in social networks is more detailed than in GPS trajectories, since the user has to register with some personal information, while in GPS traces there is only the object identifier and the spatio-temporal points. As a consequence, it is more challenging to extract socio-demographic profiles from GPS traces, where no personal information is available.

Our work is different from the previous ones, since we propose a socio-demographic profile model, which contains a set of features and rules that describe the behavior of a profile. For instance, an object with student profile goes frequently to a school for long periods of time. We also propose a *moving object history model*, which summarizes the *individual* user movement history in a way that it can be matched with the profile model. As a result, we give the similarity of a user with a given socio-demographic profile or to multiple profiles.

Table 2 summarizes the main differences of existing approaches and our proposal. It is important to point out that, to the best of our knowledge, T-Profiles is the first attempt to extract socio-demographic profiles from GPS trajectory data.

Table 2 – Summary of Related Works.

Works/Features	Type of data	Socio-Dem profiles	Define a Profile Model	Individual History Models	Similarity of trajectories and profiles	Algorithm for trajectory socio profiles
(FURLETTI et al., 2013)	GSM calls					
(BAYIR; DEMIRBAS; EAGLE, 2010)	GSM calls					
(YING et al., 2010)	GPS					
(XIAO et al., 2010)	GPS					
(HUNG; CHANG; PENG, 2009)	GPS			X		
(TRASARTI et al., 2011)	GPS			X		
(CHEN; PANG; XUE, 2014)	GPS			X		
(BAGLIONI et al., 2009)	GPS	X	X			
(ZHENG; ZHENG; YANG, 2009)	GPS + Social Net	X		X	X	
(YE et al., 2009)	GPS			X		
(BI et al., 2013)	Logs+ Social Net	X				
(MACKINNON; WARREN, 2007)	Social Net	X				
(JIANG; FERREIRA; GONZÁLEZ, 2012)	Census data	X				
T-Profiles	GPS	X	X	X	X	X

3 INFERRING SOCIO-DEMOGRAPHIC PROFILES FROM TRAJECTORIES

In this chapter we present a set of definitions to infer socio-demographic profiles from moving object trajectories. First, we propose a profile model, with a set of rules that a moving object should satisfy to belong to a profile. This model is presented in Section 3.1. Second, we define a moving object history model, in order to allow a matching between a profile model and the history model. The history model is presented in Section 3.2. Third, in Section 3.3, we present the similarity measures for matching the *moving object history model* with each *profile model*, in order to obtain a set of socio-demographic profiles from moving trajectories. Forth, in Section 3.4, we present in details the proposed method T-Profiles which performs the match between *profile models* and the *moving object history model* using the presented similarity functions.

Figure 2 gives an overview of our proposal. Taking as input semantic trajectories (ALVARES et al., 2007), we first compute the trajectory history model. Then, we compute the similarity between the *history model* and the profile model. The output is a set of trajectories labeled with one or more profile names.

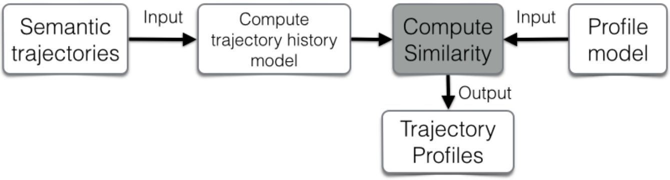

Figure 2 – Overview of the proposal.

Before defining the profile model and the history model, there are some basic concepts that need to be introduced. The approach of this work is to extract the socio-demographic profiles from moving object trajectories. The trajectories are generated by a mobile device, that records the location of a moving object during a period of time. Definition 1 contains the formal definition of trajectory, which in this work we call raw trajectory.

Definition 1 (**Raw Trajectory**)**.** A raw trajectory T is a set $\langle tid, p_0, p_1, ..., p_n\rangle$ where tid is the trajectory identifier and $p_i = (x_i, y_i, t_i)$ is a point, being x_i and y_i the spatial coordinates which correspond to the location where the object was present during the instant of time t_i, for $i = 0, ..., n$ and $t_0 < t_1 < ... < t_n$.

A raw trajectory does not have any semantic information available for analysis. Therefore, in this work we consider semantic trajectories. Several definitions can be found in the literature for semantic trajectories (BOGORNY et al., 2014) (PARENT et al., 2013), but for the sake of simplicity, we consider as semantic trajectory a set of important places called *stops*, as originally introduced in (SPACCAPIETRA et al., 2008) and extended in (ALVARES et al., 2007).

A semantic trajectory A is a sequence of stops $\langle stop_1, ..., stop_i\rangle$ ordered in time. Figure 1 shows an example of a semantic trajectory A that has four stops $\langle Home, University, ShoppingMall, Bar\rangle$. The computation of semantic trajectories as a set of stops and moves is a trivial step, once several algorithms have been proposed to compute stops from trajectories, such as (ALVARES et al., 2007), (PALMA et al., 2008), (ROCHA et al., 2010), (ARBOLEDA et al., 2014), (MORENO; PATINO; BOGORNY, 2015). More details about how to compute stops can be found in (BRAZ; BOGORNY, 2012), that presents a summary of the main methods in the literature. For the sake of simplicity, we assume that each stop is associated to a POIType and has a start and end time. The formal definition of stop is given in Definition 2.

Definition 2 (**Stop**)**.** Let *POIType* be a type of Point of Interest (POI), *startTime* and *endTime* be the start and end time that delimit the interval $[startTime, endTime]$ in which a moving object *oid* stays at a POI of *POIType*. Then, a stop is a tuple $(oid, POIType, startTime, endTime)$.

Having defined semantic trajectories and stops, in the following section we propose a profile model with features that can be obtained from semantic trajectories.

3.1 PROFILE MODELING

A profile is a set of features that represent a group of people with similar characteristics. Here we focus on a specific type, the sociodemographic profiles, where the set of features describe a given sociodemographic category of people, such as student, worker, retired, etc.

Every profile has a set of characteristics which describe a profile/category. For example, the features *go to school, four or five times a week* describe a *student* profile. *Go to work, five times a week*, describe a *worker* profile. These examples of profiles are not mutually exclusive, since a worker can also be a student.

In order to extract socio-demographic profiles from moving object trajectories we must provide a profile model with features that can be extracted and compared to moving object semantic trajectories. For this purpose, we assume that four main features describe a socio-demographic profile: the *type of place* where people go (called POIType), *when* they go, *how often* and for *how long* they stay there. With this set of features we define a *profile rule.*

Definition 3 (**Profile rule for a POIType**). Let $POIType$ be a type of POI and $p \in \mathcal{P}$ be a profile name. Then a profile rule r for a $POIType$ and a profile p is a tuple:

$$r = (p, POIType, freq, \omega_f, timeU, weekPeriod, dayPeriod, duration, \omega_d),$$

where $freq$ is the frequency that a $POIType$ is visited in a time unit $timeU$, during certain periods of the day $dayPeriod$, and period of the week ("weekday", "weekend"or "week") $weekPeriod$, $duration$ is an interval that describes the expected amount of time spent at $POIType$ in the specified period of the day and week. ω_f and ω_d are the weights for the attributes $freq$ and $duration$, respectively, which $\omega_f, \omega_d \in [0, 1]$ and $\omega_f + \omega_d = 1$.

Table 3 shows some examples of profile rules for different POITypes (Workplace, School, Restaurant), for the profile names *Full time worker*, *Part time worker*, *Student* and *Retired.* For the first two profile names, *Full time worker* and *Part time worker*, the rule expresses that the POIType *Workplace* must be visited at least 4 times a week. The main difference between a full time and part time worker is the *duration.* In order to distinguish these profiles we define a weight to express that an attribute is more important than another one. Therefore, the attributes $freq$ and $duration$ are accompanied by a weight ω that indicates the importance of the attribute to a specific rule. In the examples of *Full time* and *Part time worker*, the *duration* has a higher weight (0.7) in the rule than the frequency (0.3).

For some profiles, a rule can express that a specific POIType should not be visited. For instance, a *Retired* should not visit a Workplace. To support this type of profile rule, we allow the definition of positive and negative rules, which are expressed through the attribute

Table 3 – Examples of Profile Rules

Profile Name p	$POIType$	$freq$ (ω_f)	$timeU$	$weekPeriod$	$dayPeriod$	$duration$ (ω_d)
Full time worker	Workplace	4 (0.3)	week	NA	NA	07:00 - 09:00 (0.7)
Part time worker	Workplace	4 (0.3)	week	NA	NA	03:00 - 05:00 (0.7)
Student	School	4 (0.5)	week	weekDay	NA	03:00 - 07:00 (0.5)
Retired	Workplace	0 (1)	NA	NA	NA	NA
Retired	Restaurant	1 (1)	week	NA	NA	NA

frequency. For positive rules the frequency attribute should be above zero ($freq > 0$), and for negative rules $freq = 0$. Table 3 shows one negative rule for the profile name *Retired*, where the frequency is zero for POIType *Workplace* and weight equal to 1. Notice that *Retired* also has one positive rule. A *Retired* is supposed to go once a week to a POIType *Restaurant*. The negative rules will have a special treatment in the matching process as explained in Section 3.3.

In case any of the attributes $weekPeriod$, $dayPeriod$ or $duration$ are not considered relevant, they can be set as Not Applicable (NA). The only exception is the attribute $timeU$, which can only assume NA when $freq = 0$, i.e., when the profile rule is negative.

The rules can be defined for more general profiles, such as no occupation, worker, and student, or in a more specialized level, such as part time worker, full time worker, etc. Figure 3 shows an example of different levels of profile categories which can be defined in the profile model. We emphasize that the hierarchy shown in Figure 3 is open for modifications depending on the necessities of the application domain, since it represents only an example.

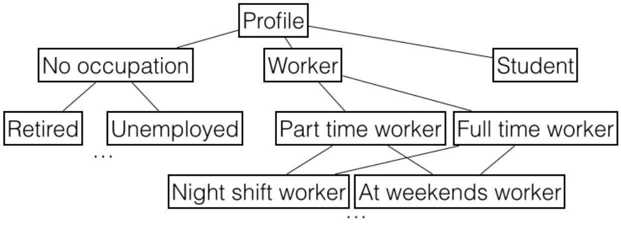

Figure 3 – Example of profile hierarchy.

Having defined the set of rules for a POIType it is possible to define the profile model as follows.

Definition 4 (**Profile model**)**.** Let $p \in \mathcal{P}$ be a profile name, a profile model for p, called $\mathcal{R}^p$, is a set of profile rules for POITypes associated with the profile name p.

Once the profile model is defined, we introduce a *moving object history model*, which summarizes the trajectories of moving objects. This model and the matching process are introduced in the following sections.

3.2 MOVING OBJECT HISTORY MODELING

The set of all stops of a moving object characterize its *movement history*. This history corresponds to the whole period that the object was tracked (e.g. one week, one month), i.e., the mobility diary. Definition 5 formalizes the object history extracted from semantic trajectories.

Definition 5 (**Object History**). An object history $h = \langle stop_1, \ldots, stop_n \rangle$ is the sequence of stops belonging to the same object such that

$$\forall i \in \{1, ..., n-1\}, endTime_i \leq startTime_{i+1}$$

where $endTime_i$ and $startTime_i$ refer to the *endTime* and *startTime* of the i-th stop of the sequence, respectively.

Having defined the object history and its properties, we are able to define a *moving object history model*. The goal is to summarize the information of an object history to a structure as similar as possible to the profile model.

Definition 6 (**Moving Object History Model**). Let *oid* be a moving object identifier and h be its trajectory history. Then a *moving object history model* $\mathcal{M}_h$ for the object history h, is a set of tuples $m \in \mathcal{M}_h$:

$$m = (oid, POIType, avgFreq, weekPeriod, dayPeriod, avgDuration)$$

where *POIType* is a type of POI, *avgFreq* is the average frequency that *oid* visits *POIType*, *weekPeriod* specifies when this happens (weekdays, weekends or whole week), *dayPeriod* indicates the period of the day (morning, afternoon, evening, night) that *oid* visits *POIType*, and *avgDuration* is the average amount of time that the object spends at *POIType* at that weekPeriod and dayPeriod. All these values are extracted from the object history h.

Each behavior tuple $m \in \mathcal{M}_h$ represents the summary of a subset of stops from the object history h with the same *POIType* for a *weekPeriod* (weekday, weekend and whole week) and *dayPeriod.* Table 4 shows an example of a *moving object history model.*

Table 4 – Example of Moving Object History Model

oid	*POIType*	*avgFreq*	*weekPeriod*	*dayPeriod*	*avgDuration*
1	Workplace	0.71	weekday	Afternoon	04:15
1	Workplace	0.71	week	Afternoon	04:15
1	Restaurant	0.14	weekend	Afternoon	01:50
1	Restaurant	0.14	week	Afternoon	01:50
1	Supermarket	0.14	weekend	Morning	02:42
1	Supermarket	0.28	weekday	Evening	00:25
1	Supermarket	0.14	week	Morning	02:42
1	Supermarket	0.28	week	Evening	00:25
1	...	...	...	...	...

It is important to notice that, since the history model is independent of the profile model, we compute the *avgFreq* and *avgDuration* of the history based only on the subset of stops in the history, not considering the profile model at this point. The *avgFreq* is computed as according to equation (3.1), where we divide the number of days the object has visited a POIType during *weekPeriod* and *dayPeriod*, represented by *visits*, by the total number of days of the history $days(h)$. This way, we get the distribution of visits through the days present in the object history. The computation of *avgFreq* is necessary for summarizing the distribution of visits to a POIType through time, i.e. the frequency of visits regarding a certain period of time.

$$avgFreq = \frac{visits}{days(h)} \tag{3.1}$$

For the attribute *avgDuration*, we use the average of the durations of the stops with the same *POIType*, *weekPeriod* and *dayPeriod.* In the following section we present a set of similarity measures for matching the profile model with the user history model.

3.3 MOVING OBJECT HISTORY MODEL AND PROFILE MODEL MATCHING

Having defined the moving object history model it is possible to verify the similarity of this model with a profile model. As defined in section 3.1, a profile model can have two types of profile rules: positive and negative. For each type of rule the matching process is different. In Equation (3.2) we give the function that computes the similarity between a positive profile rule r and a tuple m of the moving object history model $\mathcal{M}_h$. It represents the sum of the similarities of frequency (sim_f) and average duration (sim_d) multiplied by their corresponding weights (ω_f and ω_d). The tuple m, in order to be considered for matching, should have the same POIType, weekPeriod, and dayPeriod of the ones in the profile rule r, such that $POIType_m = POIType_r$, $weekPeriod_m = weekPeriod_r$, and $dayPeriod_m = dayPeriod_r$.

$$sim_{pos}(m, r) = sim_f \cdot \omega_f + sim_d \cdot \omega_d \tag{3.2}$$

The similarity for frequency sim_f and duration sim_d are defined by functions that follow the same idea of a set membership function in fuzzy logic (ZADEH, 1965).

The frequency and the duration similarity can be implemented in many different ways. After performing several experiments over real data, we implemented the following in T-Profiles: the similarity for frequency (sim_f) is defined by equation (3.3) and illustrated in Figure 4, where $avgFreq_m$ is the average frequency computed in the history model tuple m, $freq_r$ and $timeU_r$ are respectively, the frequency and the time unit defined in the profile rule r. The function $days(timeU_r)$ returns the number of days that the time unit represents (e.g. if $timeU_r = week$, then $days(timeU_r)$ returns 7).

$$sim_f = \begin{cases} 0 & \text{if } avgFreq_m = 0 \\ \frac{avgFreq_m \cdot days(timeU_r)}{freq_r} & \text{if } avgFreq_m < \frac{freq_r}{days(timeU_r)} \\ 1 & \text{if } avgFreq_m \geq \frac{freq_r}{days(timeU_r)} \end{cases} \tag{3.3}$$

Figure 4 shows an example of sim_f where $freq_r = 4$ and $timeU_r = 7$. If $avgFreq_m$ is between the interval [0, 4] the similarity increases linearly from 0 to 4. For $avgFreq_m \geq \frac{freq_r}{days(timeU_r)}$ the $sim_f = 1$.

For instance, considering the values defined in Figure 4, where $freq_r = 4$ and $timeU_r = 7$, then a history model with $avgFreq = 5$ visits per week has $sim_f = 1.0$, while a history model with $avgFreq = 2$ visits per week has $sim_f = 0.5$.

The duration similarity (sim_d) is defined by equation (3.4), and is illustrated in Figure 5, where $duration_r$ is defined as the interval [1:00, 2:00]. As can be seen in Figure 5, an $avgDuration_m$ between 1 and 2 hours will have $sim_d = 1$. For an $avgDuration_m$ between 0.5 hour and 1 hour the similarity increases linearly from 0 to 1, and for an $avgDuration_m$ between 2 hours and 2.5 hours the similarity decreases linearly from 1 to 0.

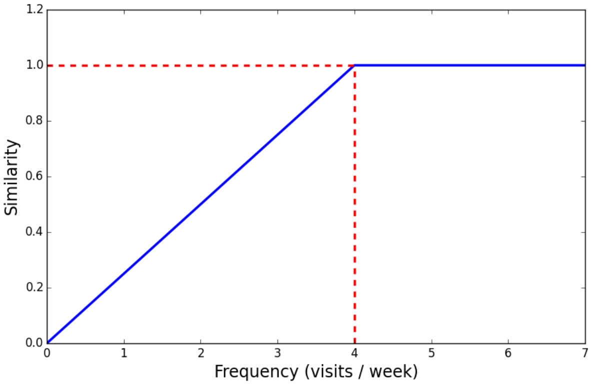

Figure 4 – Frequency similarity function sim_f for $freq_r = 4$ and $timeU_r = 7$.

$$sim_d = \begin{cases} 0 & \text{if } avgDuration_m < minGlobal \vee \\ & \quad avgDuration_m > maxGlobal \\ 1 - \frac{minDur - avgDuration_m}{minDur - minGlobal} & \text{if } minGlobal < avgDuration_m \wedge \\ & \quad avgDuration_m < minDur \\ 1 & \text{if } minDur \leq avgDuration_m \wedge \\ & \quad avgDuration_m \leq maxDur \\ 1 + \frac{maxDur - avgDuration_m}{maxGlobal - maxDur} & \text{if } maxGlobal > avgDuration_m \wedge \\ & \quad avgDuration_m > maxDur \end{cases} \tag{3.4}$$

where
$duration_r = [minDur, maxDur]$
$minGlobal = minDur - (minDur * 0.5)$
$maxGlobal = maxDur + (minDur * 0.5)$

For instance, considering the values $minDur = 1.0$ and $maxDur = 2.0$ defined in Figure 5, a history model with $avgDuration = 1.0h$ would have $sim_d = 1.0$ while one with $avgDuration = 0.9h$, few minutes less than the earlier example, would have $sim_d = 0.8$.

We define the $minGlobal$ and $maxGlobal$ limits according to the $minDur$, where the similarity function sim_d uses 0.5 of the $minDur$ to

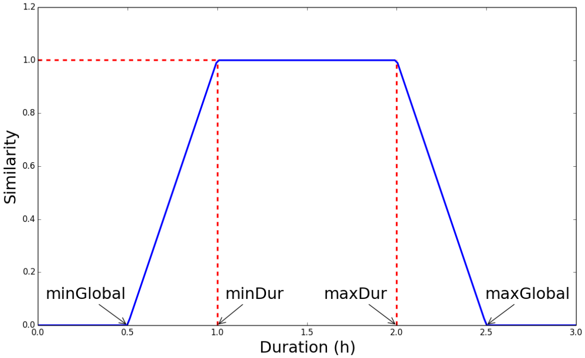

Figure 5 – Duration similarity function sim_d for $duration_r$ = [1:00 - 2:00].

define the limits of the function. We decided to use 0.5 of the $minDur$ as a way to keep the similarity function flexible for different scales. The purpose of using $minGlobal$ and $maxGlobal$ is to obtain some similarity between profile models and object history even if the duration of the visits is not exactly inside the $minDur$ and $maxDur$ interval defined in the rule. Many objects with visits in the history having duration a little bellow or above the minimal and maximal duration defined in the profile rules, would have no similarity without defining the $minGlobal$ and $maxGlobal$ limits. Using the outer limits $minGlobal$ and $maxGlobal$, we distantiate from a binary approach to a more continuous one for the similarity of the duration sim_d.

The similarity for negative rules is defined by equation (3.5), where if the $POIType$ defined in a profile rule r is not present in the moving object history model $\mathcal{M}_h$ it has $sim = 1$. It means that the moving object did not visit the $POIType$, thus fulfilling the negative behavior implied by the rule.

$$sim_{neg}(\mathcal{M}_h, r) = \begin{cases} 1 & \text{if } POIType_r \notin \mathcal{M}_h \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The total similarity between a *moving object history model* $\mathcal{M}_h$ and a profile name $p \in \mathcal{P}$ is given by the function $MATCH$ in Equation (3.6). In general words, it is the sum of the similarities of all the profile rules defined in $\mathcal{R}^p$, positives and negatives, represented by sim_{pos} and

sim_{neg}, respectively, divided by the total number of rules of that profile name p.

$$MATCH(\mathcal{M}_h, p) = \frac{\sum sim_{pos}(m_i, r_j) + \sum sim_{neg}(\mathcal{M}_h, r_k)}{|\mathcal{R}^p|} \tag{3.6}$$

where $\mathcal{R}^p$ is the set of rules of the profile name p.

Having defined the matching process for the profile model and the moving object history model, in the following section we present the algorithm T-Profiles, which extracts socio-demographic profiles from trajectory data.

3.4 T-PROFILES: AN ALGORITHM FOR DISCOVERING TRAJECTORY PROFILES

Earlier in this Chapter 3 we presented a general view of the proposed method in Figure 2. In this section, we present more details about T-Profiles.

Listing 3.1 shows the pseudo-code of the proposed algorithm to extract profiles from trajectory data, named T-Profiles. The algorithm receives as input a set of semantic trajectories T, a profile model $\mathcal{R}$, and the minimal similarity degree ϵ for an object to be considered similar to a profile name. The output is a set of moving objects labeled with a profile name p and the degree of similarity between the object history and the profile.

The first step is to compute the history model for each moving object in T (lines 11, 12) (detailed in Listing 3.2). Having computed the moving object history model, it will be compared to all rules of each profile name p in the profile model $\mathcal{R}^p$ (lines 13-37).

In the lines 17 and 23 are the similarity functions used, which are detailed in Section 3.3. These values are used to compute the matching between a moving object history model and a profile name (line 32).

If one or more of the negative rules of a profile name are not satisfied by the object history model, the similarity is set to zero, since the negative rules are mandatory (line 30). We consider that a moving object has a history similar to a profile name only when the similarity is above a given threshold ϵ, so if the total similarity is greater than the threshold ϵ, then the moving object identifier, the profile name, and the similarity degree are added to the output set ψ of trajectory profiles (line 35). This step finishes the analysis of one profile name and the algorithm returns to line 13 to test the next profile name in $\mathcal{P}$ with the current object history. Notice that the algorithm has the capability to return multiple profiles, i.e., a moving object can belong to several profiles in case the match is above ϵ.

Listing 3.1 – Pseudo-code of the algorithm T-Profiles

```
1   Input:  $T$ // set of semantic Trajectories
2           $\mathcal{R}$ // profile model = $R^p, \forall p \in \mathcal{P}$
3           $\epsilon$ // minimal similarity degree for
4             // an object belong to a profile
5
6   Output: $\psi$ // set of moving object profiles
7
8   Method:
9
10        $\psi = \{\}$ //empty set
11        for each moving object history $h \in T$ do
12          $\mathcal{M}_h$ = buildMovingObjectHistoryModel($h$)
13          for each profile name $p \in \mathcal{P}$ do
14            sumPos = 0
15            for each positive rule $r \in \mathcal{R}^p$ do
16              for each $m \in \mathcal{M}_h$ do
17                sumPos = sumPos + $sim_{pos}(m, r)$
18              end for
19            end for
20            negativeRulesNotHold = False
21            sumNeg = 0
22            for each negative rule $r \in \mathcal{R}^p$ do
23              aux = $sim_{neg}(\mathcal{M}_h, r)$
24              sumNeg = sumNeg + aux
25              if aux = 0
26                negativeRulesNotHold = True
27              end if
28            end for
29            if negativeRulesNotHold
30              MATCH = 0.0
31            else
32              MATCH = (sumPos + sumNeg) / |$\mathcal{R}^p$|
33            end if
34            if MATCH > $\epsilon$
35              $\psi$.add($oid$, $p$, MATCH)
36            end if
37          end for
38        end for
39        return $\psi$
```

Listing 3.2 shows the procedure to build the moving object history model. It receives as input an object history h and outputs a moving object history model $\mathcal{M}_h$. For each combination of $POIType$, week period and day period in the object history, the function *buildHistoryTuple()* builds a history model tuple, as shown in Section 3.2 (Definition 6) (line 8), which is added to the history model (line 10). The function $days(h)$ (line 8) returns the number of days in the object history, which is used to compute the avgFreq in the moving object history model. Finally, in line 15, the algorithm returns the *moving object history model* $\mathcal{M}_h$ built from the object history.

Listing 3.2 – buildMovingObjectHistoryModel

```
1  Input:  $h$ // object history
2  Output: $\mathcal{M}_h$ // moving object behavior model
3  Method:
4  $\mathcal{M}_h = \{\}$
5  for each $POIType \in h$ do
6    for each weakPeriod $w$ do
7      for each dayPeriod $d$ do
8        tuple = buildHistoryTuple($oid, POIType, w, d, days(h)$)
9        if tuple $\neq \phi$
10         $\mathcal{M}_h$.add(tuple)
11       end if
12     end for
13   end for
14 end for
15 return  $\mathcal{M}_h$
```

4 EXPERIMENTAL EVALUATION

In this chapter we evaluate our proposal using three datasets. The first one is a census dataset from Italy where we have the ground truth. This experiment is explained in section 4.1. The second experiment considers GPS trajectory data generated from car trajectories in Tuscany, explained in Section 4.2, and the third experiment considers GPS data from the city of Florence, explained in Section 4.3. All these data are obtained from a collaboration Brasil/Italy in the context of the SEEK project[1].

Before we go into details, we have to introduce some well-known classification concepts that are used to evaluate the results. A classification technique is a systematic approach to build classification models from an input dataset. These techniques are learning algorithms that given a training set, build a model that links a set of attributes to a label (or class). This model must be able to adapt well to the training data, predicting the class of future objects that it does not know, i.e., objects that are not present in the training set.

The performance of a classification model is based on the number of test objects that are classified correct and incorrectly. The test objects are those that the labels are known, but they were not used in the training step. These objects can be organized in a table called *confusion matrix*, where the lines represent the real classes and the columns represent the predicted classes. Good results correspond to large numbers in the main diagonal of the matrix and small ones in others positions (TAN; STEINBACH; KUMAR, 2005).

To facilitate the understanding of the performance of a classification model, we use the concepts of precision, recall and F-measure, which summarize some characteristics of the results. Precision is a metric that indicates how many of the objects in a class are classified correctly to the same class. Recall is a metric that indicates how many of the objects classified into a class really belong to that class. F-measure is the average between precision and recall. Precision, recall and F-measure formulas are shown in equation (4.1), where given a class c as example, tp are the objects that belong to class c and were classified with the right labels (true positive), fp are the objects that belong to the class c but were classified in a different one (false positive) and fn are the objects that do not belong to class c but were classified into it (false negative) (TAN; STEINBACH; KUMAR, 2005).

[1] http://www.seek-project.eu/

$$\begin{aligned} precision &= \frac{tp}{tp + fp} \\ recall &= \frac{tp}{tp + fn} \\ F - measure &= \frac{precision + recall}{2} \end{aligned} \tag{4.1}$$

As baseline, we compare our proposal with existing classification algorithms, RIPPER and C4.5. RIPPER is a rule learning algorithm that builds rules based on labeled examples given in the training step (COHEN, 1995). C4.5 is an algorithm used to generate decision trees which can be used for classification of examples through supervised learning (using labeled examples) (QUINLAN, 1993). We use the implementation of these algorithms available in Weka (FRANK et al., 2005), which are JRIP (RIPPER) and J48 (C4.5).

We do not compare the results of our method with approaches for trajectory profiles because to the best of our knowledge there is no approach that infers socio-demographic profiles from trajectories. The method was developed using *Java* as programming language and the datasets were stored in *Postgres* with the *PostGIS* extension.

4.1 CENSUS TRAJECTORIES

As it is still very difficult to obtain a dataset of semantic trajectories with a ground truth, we first evaluate the algorithm T-Profiles on a "trajectory" dataset generated from census data, where we have the ground truth. This dataset is a census of activity diaries collected in Italy during the year of 2008 which contains 40944 participants, having the socio-demographic profile of each individual that was interviewed. Each activity diary corresponds to the activities of one person during one day, and can be interpreted as the "semantic trajectories" of each individual, because they contain the place of activity (that corresponds to the POIType of the stops), the activities performed at the place, and the begin and end time of the activities. Examples are: sleeping at home from 10PM to 8AM, profile *retired*; working at a workplace from 10AM to 5PM, profile *worker*; studying at the university from 2PM to 6PM, profile *student*, etc. The most significant profiles in the database are: *worker*, *retired*, *unemployed*, *housewife with kids* and *student*.

As one day of activities is not enough to determine the profile of a person, we preprocessed the data, grouping diaries that belong to the same socio-demographic profile from the census data, considering 14 days of activities, taking into account the days of the week, and considering that the last POIType of one day should be the same as the first POIType of the next day. As a result, we obtained trajectories of 14 days long for 829 objects. We chose 14 days for the history length because it englobes 2 weeks. We assume that if a behavior is performed during 1 week and repeated in the next week, the behavior is a frequent pattern. The amount of objects for each profile is shown in Figure 6.

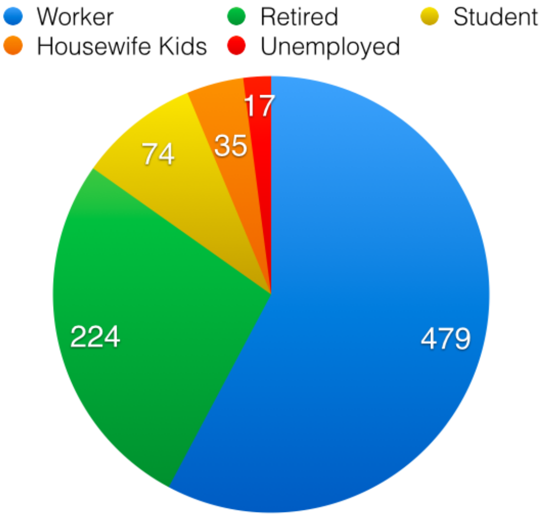

Figure 6 – Number of 14-day-long object traces obtained from the census data diaries.

Table 5 shows the rules considered in this experiment. A *worker* is identified by the POIType Workplace, that should be visited with a frequency of 4 times a week with duration between 3 and 9 hours. We define a broad range for duration to obtain all types of workers (full time and part time). Notice that we defined a higher weight for the frequency (0.8), because this attribute is more important than the duration.

The rule for the profile named *Student* expresses that this profile should visit a POIType related to educational institutions, such as schools or universities, for at least 3 times a week on weekdays, with a duration between 3 and 6 hours per day, to include full time and part

Table 5 – Profile models for profiles related to occupation status.

Profile Name p	$POIType$	$freq$ (ω_f)	$timeUnit$	$weekPeriod$	$dayPeriod$	$duration$ (ω_d)
Worker	Workplace	4 (0.8)	week	NA	NA	03:00 - 09:00 (0.2)
Student	School/Univ.	3 (0.5)	week	weekday	NA	03:00 - 06:00 (0.5)
Retired	Workplace	0 (1)	NA	NA	NA	NA
Retired	School/Univ.	0 (1)	NA	NA	NA	NA
Retired	Bar	1 (1)	week	NA	Morning, Afternoon	NA
Unemployed	Workplace	0 (1)	NA	NA	NA	NA
Unemployed	School/Univ.	0 (1)	NA	NA	NA	NA
Unemployed	Bar	1 (1)	week	NA	Evening	NA
Unemployed	Restaurant	1 (1)	month	NA	NA	NA
Unemployed	Gym / Sport court	2 (1)	month	NA	NA	NA
Housewife Kids	School/Univ.	3 (0.5)	month	weekday	NA	00:20 - 01:00 (0.5)
Housewife Kids	Commercial estab.	3 (1)	week	weekday	NA	NA

time students. The weights for the attributes $freq$ and $duration$ are both 0.5.

For the profile *Retired*, we have defined two negative rules related to workplace and educational places, to distinguish between workers and students, since it is expected that most retired do not have a workplace and do not go to school. However, these rules are not enough to distinguish a retired from an unemployed. Then, as they are supposed to go more often to bars or cafes, we create a rule with this kind of POIType, in the period of morning and afternoon, i.e., during the day.

The profile *Unemployed* may have similar behavior to the *Retired*, having no working place and not going to school to distinguish these profiles. To distinguish an unemployed from a retired we define three positive rules: POIType Bar visited during the evening, POIType Restaurant visited only once a month, and visits to sport places. One can complain that an unemployed could study, but in this case it should belong to the profile student.

A housewife that has children can be identified if the person visits educational places such as schools. But the difference from the profile student is the frequency and the duration. The profile does not need to go every day to take the child to the school, but should at least visit a POIType school sometimes to express that there is a relationship with educational place. Defining a rule forcing a housewife with kids to go very frequently to educational places would limit the discovery only of cases where the housewife takes the kids to school everyday.

We performed experiments considering three different values for the similarity threshold ϵ between the profile model and the history model: 60%, 70% and 80%. Table 6 shows the results for similarity ϵ of 60%, 70% and 80%. For similarity 70%, for instance, T-Profiles detected 478 workers of 479, and 73 of 74 students. For the profile Housewife Kids, 24 instances were discovered. The most difficult classification is to distinguish unemployed and retired, because their behavior are very similar, but still 158 retired and 9 unemployed were detected.

Table 6 – Profiles for 60%, 70% and 80% similarity.

Profiles	Total	$\epsilon = 60\%$	$\epsilon = 70\%$	$\epsilon = 80\%$
Worker	479	479	478	473
Housewife Kids	35	28	24	20
Unemployed	17	9	9	9
Retired	224	185	158	158
Student	74	74	73	72

We show the confusion matrix for the threshold 70% in Table 7. In this table, each row corresponds to a profile name and each column represents the number of objects classified by T-Profiles in that profile name. The column *Inconclusive* contains all objects that did not reach the minimal similarity with any profile name in the profile model. From the total of 479 workers, 478 are correctly classified, and only one is inconclusive. Similarly, for the profile Student, from the total of 74, 73 are correctly classified and only 1 is inconclusive. For the profile Housewife Kids, 23 instances were correctly classified, 8 cases are inconclusive and 3 were wrongly classified as Retired and 1 as Unemployed. 167 retired are correctly classified, 47 remain inconclusive, 5 are wrongly classified as Housewife Kids, and 4 as Unemployed.

Table 7 – Confusion matrix using $\epsilon = 70\%$

	Inconclusive	Worker	Housewife Kids	Unemployed	Retired	Student
Worker	1	478	0	0	0	0
Housewife Kids	7	0	24	1	3	0
Unemployed	2	0	2	9	4	0
Retired	53	0	6	7	158	0
Student	1	0	0	0	0	73

Table 8 shows the precision and recall, considering the similarities for each profile name as well as the average for all objects. As can be seen in this table, the results are not too different as the similarity threshold changes. T-Profiles shows a very high average precision, about 97%. The recall is also high, between 89% and 93% with these values of ϵ.

Table 8 – Precision and Recall

	$\epsilon = 60\%$		$\epsilon = 70\%$		$\epsilon = 80\%$	
	Precision	Recall	Precision	Recall	Precision	Recall
Worker	1.000	1.000	1.000	0.998	1.000	0.983
HousewifeKids	0.587	0.771	0.767	0.657	0.769	0.571
Unemployed	0.571	0.471	0.571	0.471	0.667	0.471
Retired	0.949	0.835	0.954	0.746	0.954	0.746
Student	1.000	0.973	1.000	0.986	1.000	0.973
Avg.	0.960	0.935	0.969	0.903	0.971	0.890
Avg. F-measure	0.946		0.936		0.921	

We point out that the precision for Worker and Student is 100% for the three different similarity thresholds ϵ, i.e., when the algorithm classifies an object as worker or student, this object is really a worker or a student.

In these results we considered only the highest similarity for each profile name, but we can also analyze all similarities that are above the threshold ϵ, having *multiple profiles*. Table 9 shows some examples of the output of T-Profiles. Each row corresponds to an object. The multiple profile column shows all profile names that have similarity above 80%. For the object 842, for instance, the similarity with Worker and Housewife Kids is above 90%, so this object is labeled as Worker and Housewife Kids.

Table 9 – Multiple profiles found using $\epsilon = 80\%$

oid	Worker	Housewife Kids	Unemployed	Retired	Student	Multiple profile
16	0.000	0.174	0.800	0.980	0.000	Retired, Unemployed
131	0.000	0.261	0.800	0.980	0.000	Retired, Unemployed
842	1.000	0.949	0.000	0.000	0.179	Worker, Housewife Kids
600	1.000	0.897	0.000	0.000	0.100	Worker, Housewife Kids

We compare our results with two classification algorithms, RIPPER (COHEN, 1995) and C4.5 (QUINLAN, 1993). We use the implementation in Weka (FRANK et al., 2005) of these algorithms, JRIP and J48, respectively, using 10-fold cross-validation. We use as input for these algorithms the history model generated by our method, i.e., all methods have the same input. Notice that we helped the classification algorithms giving as input the trajectories pre-processed according to our history model, that is also a contribution of T-Profiles. Table 10 shows the precision and recall for T-Profiles with $\epsilon = 70\%$, RIPPER, and C4.5.

The traditional classification methods classify every instance of the data. As a consequence, in general, they also classify incorrectly more instances. This can be noticed in Table 10. The precision of T-Profiles for each profile is almost always better than both classifier algorithms. For the profile Unemployed, for instance, the precision of T-Profiles is 0.57, while it is 0.17 for RIPPER and 0.25 for C4.5.

Table 10 – Comparing T-Profiles, RIPPER and C4.5.

	T-Profiles		RIPPER		C4.5	
	Precision	Recall	Precision	Recall	Precision	Recall
Worker	1.000	0.998	0.981	0.994	1.000	1.000
Housewife Kids	0.767	0.657	0.783	0.514	0.656	0.600
Unemployed	0.571	0.471	0.167	0.059	0.250	0.118
Retired	0.954	0.746	0.877	0.951	0.894	0.942
Student	1.000	0.986	0.972	0.946	0.973	0.973
Avg.	0.969	0.903	0.927	0.938	0.939	0.947
Avg. F-measure	0.936		0.932		0.943	

In average, T-Profiles obtained a better precision than RIPPER and C4.5, being for T-Profiles 96.9% against 92.7% and 93.9% for RIPPER and C4.5 respectively. For the average recall, T-Profiles is lower than for both RIPPER and C4.5, but notice that it is not so far behind, where T-Profiles presents 90.3% of recall while RIPPER and C4.5 present 93.8% and 94.7% respectively. Considering individual profiles, T-Profiles obtains better results among the other algorithms, specially in the precision.

4.2 OCTOSCANA DATASET

The Octoscana is a small dataset with 66 trajectories collected during an average period of 13 days, that was manually annotated, although not by the users that generated the trajectories. The dataset has several gaps in the historical trajectories, and some stops are relatively long, and may characterize some errors in the data. The profiles defined in this dataset may not be 100 % correct, but they can be used as an approximate ground truth real trajectory dataset. Table 11 shows the profiles for this data, which has 42 workers, 12 unemployed, 9 retired and 3 housewife kids. For this dataset we considered exactly the same rules as in the previous experiment, i.e., the same profile model (shown in table 5 in Section 4.1), only removing the rule for visiting sportive places because this POIType was not labeled in the data.

Table 11 – Profiles labeled in Octoscana.

Profiles	**Total**
Worker	42
Unemployed	12
Retired	9
Housewife Kids	3
Total	66

We show the confusion matrix for $\epsilon = 70\%$ in Table 12. Focusing on the profile similarity of 70%, 36 workers were correctly classified, out of 42, and no one was wrongly labeled. The profiles Retired and Unemployed are very difficult to distinguish, so T-Profiles incorrectly labeled 5 Unemployed out of 12, and incorrectly labeled 4 Retired out of 9.

Using the same profile model, T-Profiles has a better average precision than RIPPER, but it was not as precise as C4.5 (see Table 13). For the profile Housewife kids, however, it has a high precision, while RIPPER and C4.5 have precision 0.

As this dataset has only 12 objects with profile Unemployed and 9 as Retired, it may be hard that such a low number of objects have the exact behavior defined in the general profile model. So for this specific dataset, changing one rule in the profile model for Retired to visit the POIType Bar with the same frequency, but restricted to the weekPeriod weekend, T-Profiles reaches an average precision of 0.890 and recall of 0.758, as shown in Table 13.

Table 12 – Confusion matrix using $\epsilon = 70\%$

	Inconclusive	Worker	Housewife Kids	Unemployed	Retired
Worker	6	36	0	0	0
Housewife Kids	1	0	1	0	1
Unemployed	1	0	0	6	5
Retired	0	0	0	4	5

Table 13 – Comparing T-Profiles, RIPPER and C4.5

	T-Profiles		RIPPER		C4.5	
	Precision	Recall	Precision	Recall	Precision	Recall
Worker	1.000	0.857	0.891	0.976	1.000	1.000
Housewife Kids	1.000	0.333	0.000	0.000	0.000	0.000
Unemployed	0.643	0.750	0.733	0.917	0.714	0.833
Retired	0.667	0.444	0.400	0.222	0.700	0.778
Avg.	0.890	0.758	0.755	0.818	0.862	0.894
Avg. F-measure	0.824		0.786		0.878	

4.3 FLORENCE DATASET

This dataset is a very noisy dataset of residents in the city of Florence. The data is a sample of 5664 raw car trajectories, collected by an insurance company during one month, but the average tracking period of one object was 10 days, i.e., the historical movement of an object is in average 10 days, but not necessarily all objects have such a long history. Indeed, it is known that in the city of Florence it is quite hard to move by car, and sometimes it is necessary to park far from the visited POI, so the probability to obtain the correct place that an object visited is quite difficult. Even with these problems we considered the dataset for our experiments, because it is a real trajectory dataset, and we can show the potential of the proposed approach when the dataset is imprecise. We started the experiments computing the *approximate stops* for these trajectories with the method SMOT (ALVARES et al., 2007), considering 10 minutes as minimal time for a stop. A stop is created when a trajectory intersects a place defined in Open Street Maps for at least 10 minutes. The POITypes were also extracted from Open Street Maps, so the stops are only defined when the trajectory intersects a place that is present in Open Street Maps. From this dataset, we selected all trajectories with more than 10 stops labeled with their POITypes and with at least 10 days history, so obtaining a set of 417 trajectories.

Here we show the flexibility of T-Profiles, where the user can choose any level of profile category analysis, from the more general to the more specialized. We are interested in Full Time Workers, Part Time Workers, Weekend Workers and Night Workers. Considering the rules defined in Table 5, we extended the set of rules with new profiles of workers. Table 14 shows the rules for workers where the duration distinguishes full time and part time workers, and the frequency, week period and day period distinguishes weekend and night workers.

Table 14 – Profile models for worker profiles.

Profile Name p	$POIType$	$freq\ (\omega_f)$	$timeUnit$	$weekPeriod$	$dayPeriod$	$duration\ (\omega_d)$
Full time worker	Workplace	4 (0.5)	week	NA	NA	07:00 - 09:00 (0.5)
Part time worker	Workplace	4 (0.5)	week	NA	NA	03:00 - 05:00 (0.5)
Weekend worker	Workplace	1 (1)	week	weekend	NA	NA
Night worker	Workplace	3 (1)	week	NA	Evening, Night	NA

Table 15 shows the result for 70%, 80% and 90% similarity. Considering $\epsilon = 80\%$, T-Profiles labeled 36 full time workers, 16 part time workers, 31 weekend workers, 21 night workers, 4 students, and 2 retired.

Table 15 – Profiles for 70%, 80% and 90% similarity.

Profiles	$\epsilon = 70\%$	$\epsilon = 80\%$	$\epsilon = 90\%$
Full time worker	56	36	32
Part time worker	26	16	9
Weekend worker	37	31	31
Night worker	24	21	15
Student	4	4	4
Retired	2	2	0
Unemployed	1	0	0
Housewife Kids	0	0	0

Table 16 shows some examples of the output of T-Profiles, with some single and multiple profiles, where the values represent the similarity of the object with the profile model. Notice that objects 1 and 2 were labeled with multiple profiles. For oid 1 the similarity degree was 100% with *Part time* and *Weekend worker*, while object 2 had similarity of 100% with *Full time* and *Night worker*. Objects 3 and 4 had similarity above 80% with the profile category *Full time* and *Night worker*, respectively. Another example is the object with oid 5 that had similarity 100% with the profile *Student*.

Table 16 – Examples of the output of T-Profiles for $\epsilon = 80\%$

oid	1	2	3	4	5	6
Housewife Kids	0.000	0.000	0.326	0.000	0.250	0.000
Unemployed	0.000	0.000	0.000	0.000	0.000	0.750
Retired	0.000	0.000	0.000	0.000	0.000	0.852
Student	0.000	0.000	0.106	0.000	1.000	0.000
Full time worker	0.603	1.000	0.977	0.743	0.000	0.000
Part time worker	1.000	0.500	0.477	0.318	0.000	0.000
Weekend worker	1.000	0.000	0.000	0.636	0.000	0.000
Night worker	0.000	1.000	0.000	0.848	0.000	0.000
Profiles	Part time worker, Weekend worker	Full time worker, Night worker	Full time worker	Night worker	Student	Retired

5 CONCLUSION AND FUTURE WORKS

Most works in GPS trajectory data analysis and mining look for general patterns such as objects that move together in flocks, visit similar places, follow the same routes every day, and so on, extracting patterns of *groups of objects* with similar behavior. In this thesis we provide a first attempt to go deeper in the analysis of moving object trajectories, analyzing *every individual object* mobility history, in order to discover the socio-demographic status of each individual. While the discovery of socio-demographic profiles is more trivial in social networks, GSM calls, and weblog data, where far more information about the individual is available in the form of chats, comments, or messages, as well as personal information about the object that must register on the net, the discovery of socio-demographic profiles from GPS trajectories is a great challenge.

In this thesis we propose a profile model as a set of simple rules that the user can express to discover any type of profile. We also introduce a moving object history model that summarizes the historical traces of moving objects, that is independent of a specific profile model. Finally, we propose a matching process that provides the similarity between a given profile name and a moving object based on his/her trajectory summary.

The main strengths of our proposal include: (i) the extraction of personal information about people from raw GPS traces, which in general have no additional information about the moving object; (ii) the inference of multiple profiles from raw GPS traces; and (iii) the use of the same profile model for different datasets, obtaining good results, although better results can be obtained with specific profile models. We have shown that even with very noisy and low quality trajectory data, T-Profiles can correctly label a number of objects in a dataset.

T-Profiles can achieve a high precision, taking as input a simple profile model. Classical data mining algorithms could be used to extract rules from the data and these rules could be used as input for T-Profiles, so improving still more our results.

The main problem to validate the proposal is the data, which are very noisy, sparse, and do not have a ground truth. Another problem is that although we were able to generate histories of 14 days long using the census dataset, each day was the pattern of a different person, in the same profile. This aggregation may cause some distortion and join people with different habits, but it was the solution found to validate the proposed approach.

The main disadvantage of the method is that the domain user must define the profile rules. On the other hand, it has the advantage that no training set is necessary to discover profiles, since labeled data are normally not available in the domain of trajectory data.

As result of this work, a paper named *A Rule-based Method for Discovering Trajectory Profiles* was published in the conference SEKE 2015 (International Conference on Software Engineering and Knowledge Engineering) (ALENCAR et al., 2015).

As future work, we intent to perform more experiments with different types of rules, and use longer object histories, since 14 days is normally a short period to discover the general habits of a population. A weak point noticed in the proposed method is the rigid negative rules, so another future work includes the extension of negative rules to a more fuzzy approach. Another future work is the use of sequence of places visited by the object.

Another future work aims to investigate a new approach, started during this thesis, that does not need user defined rules for the profiles. The approach makes use of the knowledge about places of interest found in social media as Foursquare, that provides the average price of the place, the proportion of male and female that visit the place, if there is parking space, etc. With such information about the place (POI), it is possible to analyze the visits of users from their trajectories and infer more detailed profiles such as economical status, gender, age, etc.

REFERENCES

ALENCAR, L. A. de et al. A rule-based method for discovering trajectory profiles. In: **The 27th International Conference on Software Engineering and Knowledge Engineering, SEKE 2015, Wyndham Pittsburgh University Center, Pittsburgh, PA, USA, July 6-8, 2015**. [S.l.: s.n.], 2015. p. 244–249.

ALVARES, L. O. et al. A model for enriching trajectories with semantic geographical information. In: ACM. **Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems**. [S.l.], 2007.

ARBOLEDA, F. J. M. et al. Smot+: Extending the smot algorithm for discovering stops in nested sites. **Computing and Informatics**, v. 33, n. 2, p. 327–342, 2014. Disponível em: <http://www.cai.sk/ojs/index.php/cai/article/view/1130>.

BAGLIONI, M. et al. Towards semantic interpretation of movement behavior. In: **Advances in GIScience**. [S.l.]: Springer, 2009. p. 271–288.

BAYIR, M. A.; DEMIRBAS, M.; EAGLE, N. Mobility profiler: A framework for discovering mobility profiles of cell phone users. **Pervasive and Mobile Computing**, v. 6, n. 4, p. 435 – 454, 2010. ISSN 1574-1192. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574119210000076>.

BI, B. et al. Inferring the demographics of search users: Social data meets search queries. In: **Proceedings of the 22Nd International Conference on World Wide Web**. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. (WWW '13), p. 131–140. ISBN 978-1-4503-2035-1. Disponível em: <http://dl.acm.org/citation.cfm?id=2488388.2488401>.

BOGORNY, V. et al. Constant - A conceptual data model for semantic trajectories of moving objects. **T. GIS**, v. 18, n. 1, p. 66–88, 2014. Disponível em: <http://dx.doi.org/10.1111/tgis.12011>.

BRAZ, F. J.; BOGORNY, V. **Introdução a trajetórias de objetos móveis: conceitos, armazenamento e análise de dados**. Joinville, SC, BR: Editora Univille, 2012. ISBN 978-85-8209-002-2.

CAO, H.; MAMOULIS, N.; CHEUNG, D. W. Discovery of periodic patterns in spatiotemporal sequences. **Knowledge and Data Engineering, IEEE Transactions on**, IEEE, v. 19, n. 4, p. 453–467, 2007.

CHEN, X.; PANG, J.; XUE, R. Constructing and comparing user mobility profiles. **ACM Transactions on the Web (TWEB)**, ACM, v. 8, n. 4, p. 21, 2014.

COHEN, W. W. Fast effective rule induction. In: **In Proceedings of the Twelfth International Conference on Machine Learning**. [S.l.]: Morgan Kaufmann, 1995. p. 115–123.

DASH, M. et al. Home and work place prediction for urban planning using mobile network data. In: **IEEE 15th International Conference on Mobile Data Management (MDM)**. [S.l.: s.n.], 2014. v. 2, p. 37–42.

FRANK, E. et al. Weka - a machine learning workbench for data mining. In: MAIMON, O.; ROKACH, L. (Ed.). **The Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer, 2005. p. 1305–1314.

FURLETTI, B. et al. Analysis of gsm calls data for understanding user mobility behavior. In: **Big Data, 2013 IEEE International Conference on**. Santa Clara, California: [s.n.], 2013. p. 550–555.

GIANNOTTI, F. et al. Trajectory pattern mining. In: **Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2007. (KDD '07), p. 330–339. ISBN 978-1-59593-609-7. Disponível em: <http://doi.acm.org/10.1145/1281192.1281230>.

HUNG, C.; CHANG, C.; PENG, W. Mining trajectory profiles for discovering user communities. In: **Proceedings of the 2009 International Workshop on Location Based Social Networks, LBSN 2009, November 3, 2009, Seattle, Washington, USA, Proceedings**. [s.n.], 2009. p. 1–8. Disponível em: <http://doi.acm.org/10.1145/1629890.1629892>.

JIANG, S.; FERREIRA, J.; GONZÁLEZ, M. C. Clustering daily patterns of human activities in the city. **Data Mining and Knowledge Discovery**, Springer, v. 25, n. 3, p. 478–510, 2012.

LEE, J.-G. et al. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. **Proceedings of the VLDB Endowment**, VLDB Endowment, v. 1, n. 1, p. 1081–1094, 2008.

MACKINNON, I.; WARREN, R. H. Age and geographic inferences of the livejournal social network. In: **Statistical Network Analysis: Models, Issues, and New Directions**. Springer, 2007. p. 176–178. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73133-7$_1$4¿.

MORENO, F.; PATINO, H.; BOGORNY, V. Smot+ncs: An algorithm for detecting non-continuous stops. **Computing and Informatics**, 2015. ISSN 1335-9150.

PALMA, A. T. et al. A clustering-based approach for discovering interesting places in trajectories. In: ACM. **Proceedings of the 2008 ACM symposium on Applied computing**. [S.l.], 2008. p. 863–868.

PARENT, C. et al. Semantic trajectories modeling and analysis. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 45, n. 4, p. 42:1–42:32, ago. 2013. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2501654.2501656>.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

ROCHA, J. A. M. et al. Db-smot: A direction-based spatio-temporal clustering method. In: IEEE. **Intelligent systems (IS), 2010 5th IEEE international conference**. [S.l.], 2010. p. 114–119.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. **Data & knowledge engineering**, Elsevier, v. 65, n. 1, p. 126–146, 2008.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Pearson, 2005. ISBN 978-0321321367.

TRASARTI, R. et al. Mining mobility user profiles for car pooling. In: **Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, 2011. (KDD ’11), p. 1190–1198. Disponível em: <http://doi.acm.org/10.1145/2020408.2020591>.

XIAO, X. et al. Finding similar users using category-based location history. In: ACM. **Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems**. 2010. p. 442–445. Disponível em: <http://doi.acm.org/10.1145/1869790.1869857>.

YE, Y. et al. Mining individual life pattern based on location history. In: IEEE. **IEEE 10th International Conference on Mobile Data Management (MDM)**. [S.l.], 2009. p. 1–10.

YING, J. J.-C. et al. Mining user similarity from semantic trajectories. In: **Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks**. New York, NY, USA: ACM, 2010. (LBSN '10), p. 19–26. ISBN 978-1-4503-0434-4. Disponível em: <http://doi.acm.org/10.1145/1867699.1867703>.

ZADEH, L. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338 – 353, 1965. ISSN 0019-9958. Disponível em: <http://www.sciencedirect.com/science/article/pii/S001999586590241X>.

ZHENG, V. W.; ZHENG, Y.; YANG, Q. Joint learning user's activities and profiles from gps data. In: **Proceedings of the 2009 International Workshop on Location Based Social Networks**. New York, NY, USA: ACM, 2009. (LBSN '09), p. 17–20. ISBN 978-1-60558-860-5. Disponível em: <http://doi.acm.org/10.1145/1629890.1629894>.