Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data

A Case Study in Rennes, France

Mohamed K. El Mahrsi, Etienne Côme, Johanna Baro, Latifa Oukhellou
Université Paris-Est, IFSTTAR, COSYS-GRETTIA
F-77447 Marne-la-Vallée, France
mohamed-khalil.el-mahrsi@ifsttar.fr, etienne.come@ifsttar.fr,
johanna.baro@ifsttar.fr, latifa.oukhellou@ifsttar.fr

ABSTRACT

Data collected by Automated Fare Collection (AFC) systems are a valuable resource for studying the travel habits of large city inhabitants. In this paper, we present an approach to mining the temporal behavior of the passengers in a public transportation system in order to extract relevant and easily interpretable clusters. Such classification can be useful for several applications. It may help transport operators better know the demand of their customers and propose targeted incentives, services, and tools accordingly. From a city perspective, this may also help redesign and improve existing transportation policies. To achieve this objective, an additional step of analysis is required and the clustering results need to be contextualized. The spatial location of the different types of passengers is then of great interest. We propose a first step in this direction through a rough estimation of the regular passengers' "residence" location and the analysis of socioeconomic information available at a fine-grained spatial level. The approach is applied on a real dataset from the metropolitan area of Rennes (France) with four weeks of smart card data containing trips made by both bus and subway.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Clustering; H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

smart card, public transportation, clustering, topic models.

1. INTRODUCTION

Presently, Automated Fare Collection (AFC) systems are widely adopted by public transportation operators in large cities and metropolitan areas. These systems are based on contactless smart cards that passengers credit and use when boarding buses or accessing subway stations. While the main purpose of AFC systems is to manage revenu collection, they also offer the opportunity to study the day-to-day travel habits of passengers using the transit network. In fact, each transaction registered when a smart card is used contains fine-grained information not only about the passenger's identity but also about the time, boarding location, and —in some cases— the boarded bus or subway line. This makes it possible to reconstruct the detailed profiles of the passengers over long periods of time and analyze them to study various aspects such as travel behavior and regularity, multi-modality, turnover rates, etc.

Understanding the passengers' travel behavior can be useful in a variety of applications. For instance, it can be used to assess the performance of the transit network, detect irregularities (e.g. frauds, defective equipment, etc.), forecast demand with more accuracy, make service adjustments that accommodate with variations in riderships across weekdays and seasons, etc. From a city perspective, smart card data can also prove valuable to evaluate the public transportation accessibility level, detect if particular communities or areas are underserved, and react accordingly to consolidate and improve the existing transportation policies. In this context, however, smart card data do present some shortcomings: due to privacy concerns, the collected data are anonymized and personal information about the individuals (e.g. gender, age, address, income, etc.) are not available. Consequently, inferring the socioeconomic characteristics of the passengers can prove to be difficult.

In this paper, we demonstrate how smart card data can be used conjointly with socioeconomic data to understand passenger behavior in public transportation. The contributions of this work are the following:

- We study the extraction of travel patterns from smart card data. To this effect, we construct temporal passenger profiles based on boarding information and apply a generative model-based clustering approach to discover groups (or clusters) of passengers who behave similarly w.r.t. their boarding times.
- We study how passenger travel habits relate to socioeconomic characteristics. To do so, we start by assigning passengers based on their boarding information to "residential" areas that we established through a clustering of socioeconomic data of the city. Then, we inspect how socioeconomic characteristics are distributed over the passenger temporal clusters.

UrbComp'14, August 24, 2014, New York, NY, USA.

- We apply our approaches on socioeconomic data and a real smart card dataset covering four weeks of boarding validations registered in the metropolitan area of the city of Rennes (France).

The rest of this paper is organized as follows. Related work is discussed in Section 2. We present our approach to clustering smart card data in order to study passenger temporal habits in Section 3. Our attempt to pair smart card and socioeconomic data is reported in Section 4. Finally, we conclude the paper in Section 5 with general remarks and future research directions.

2. RELATED WORK

Bagchi et al. [2, 3] were among the first researchers to substantiate the potential of smart card data for transit planning. The authors discuss the advantages and shortcomings of smart card data w.r.t. their capability of capturing information about passenger trips and present case studies to illustrate how rules-based processing can be used to infer turnover rates, trip rates, and detect linked trips from the collected data. The authors also emphasize that since some information are not captured by smart card data (e.g. journey purpose, satisfaction w.r.t. the transport service, etc.), the latter cannot entirely supersede existing data collection approaches (mainly direct surveys) but rather complement them.

Morency et al. [16] conduct an analysis of the variability of travel behaviors in public transit based on bus trip data. Trips are aggregated into transactions, each representing the daily profile of a given smart card on a given date. A transaction is represented using 27 attributes: a smart card id, a date, a type of day (e.g. D1-Monday, D2-Tuesday, etc.), and a feature per hour indicating if the smart card was used for validation or not (e.g. feature H10=1 indicates that the smart card was validated at least once between 10:00 and 10:59). First, the authors conduct analyses in order to calculate global indicators of regularity based on the activity rate (i.e. ratio between the number of days where a smart card was validated for boarding and the number of days of observation), the number of boardings per day, and the number of different boarding stations. Later on, the authors focus on studying the regularity of the individual behavior of two cards (an elderly and a regular adult). Two aspects are studied: (i) the number of boardings per day and (ii) boarding time. For the latter, k-means clustering is applied in order to identify (separately for each user) clusters of similar days w.r.t. boarding times. A similar analysis of weekly travel behavior using the same data from [16] is conducted by Agard et al. [1]. Here, however, the bus trips are aggregated into weekly transactions each describing the 5 weekday activity of a passenger during a given week. A transaction comprises a smart card id, the fare type of the smart card, the week, and 20 binary features representing 5 weekdays with 4 periods (AM = 5:30 to 8:59, MI = 9:00 to 15:29, PM = 15:30 to 17:59, and SO = 18:00 and over) per day. Hierarchical Agglomerative Clustering (HAC) and k-means are applied to the transactions in order to study group behavior. For instance, the authors showcase how inspecting the composition of clusters w.r.t. fare type and the evolution of the latter over the studied weeks helps discover groups such as "typical workers" (typically traveling during AM peaks and PM peaks during weekdays) and atypical behaviors (e.g. a shift of regularity in students' behavior due to spring-break).

Lathia et al. [12] discuss how smart card data can be used to reveal hidden individuals' behaviors and study their response to travel incentives. First, a comparison between an online survey's results and real travel data collected from Transport for London's (TfL) Oyster smart card system is presented. This comparison attempts to characterize the difference between the perceived and the actual behavior of passengers by studying various aspects such as trips per day frequency and regularity, atypicality of travel modality and origin and destination stations, as well as cash-fare purchasing habits. Secondly, the authors demonstrate how smart card data can help study the extent to which incentives (such as time-varying fares, price capping, etc.) proposed by transport operators influence passengers' decisions.

In [21], the behavioral differences between travel card holders and Pay as you go passengers in the London Underground are studied using smart card data over a four months period. Fuse et al. [11] investigate the use of bus smart card data to determine useful information (such as travel time and number of passengers) that can be used for congestion spot analysis and the improvement of bus stops planning. In [15], a data mining approach to extracting individual transit riders' travel patterns and assessing their regularity from a large smart card dataset with incomplete information is presented. The study focuses on two aspects: (i) characterizing the spatial and temporal travel patterns of transit riders on an individual basis, and (ii) determining the regularity of these patterns. The authors use smart card data collected from Beijing's (China) bus and subway AFC systems (offering flat-rate fares and distance-based fares). The adopted methodology is composed of three steps: (i) first, each passengers transactions are retrieved; (ii) the spatiotemporal relationships between these transactions are used to generate trip chains on which (iii) data mining techniques are applied in order to extract the passenger's travel patterns and regularity information.

Inferring "community well-being" from smart card data is discussed in [13]. First, stations are mapped, based on geographic proximity, to communities and IMD (Index of Multiple Deprivation) scores obtained from national census results. Then, trip data are used to compute a station-by-station flow matrix representing locations visited by different communities. The station-to-IMD mapping and station-to-station flow matrix are used to study the correlation between IMD and flow and calculate homophily indices. Results suggest that, contrary to well-off areas attracting people from communities of varying social deprivation, deprived areas do suffer from social segregation and tend to attract people only from other deprived areas. To some extent, this work is similar to what we present in Section 4 as it involves using socioeconomic data. In [13], these are used in order to study how communities relate to each other, whereas in our context we try to study if and how they influence temporal behavior of public transportation passengers.

Personalizing transport information services based on smart card data is discussed in [14]. Among their contributions, the authors use clustering to prove that usage of public transportation can vary considerably between individuals. Each passenger's trips are aggregated into a weekday profile describing his temporal habits (four time bins are used: early morning, morning rush, day time, and evening rush) and hierarchical agglomerative clustering is used to discover groups of passengers characterizing different travel habits

(e.g. commuting regularly during both rush hours, exclusively in the evening, etc.). Contrary to the approach we present in Section 3, all weekdays are treated equally which might cause day-specific behaviors to be ignored.

Other relevant references in the field include [18, 8]. The interested reader is also referred to the survey in [22] in which the use of public transit data to improving transportation systems is discussed, as well as the thorough literature review in [19].

One key difference between the approach reported in this paper and previous work involving clustering smart card data is the nature of the involved clustering algorithm: most existing approaches to clustering smart card data (e.g. [16, 1, 8]) rely on the use of similarity-based clustering algorithms that minimize Euclidean distortions. These approaches are often criticized for yielding poor results [20]. Our approach, in contrast, relies on the use of a generative model-based clustering (mixture of unigrams). Additionally, some existing approaches consider either clustering the trips of each passenger individually [16, 15] or disregard the identities of the passengers [8, 1]. In our approach, similarly to [14], we consider each passenger (and trips pertaining to him) as a single observation and conduct the clustering accordingly.

3. CLUSTERING PASSENGERS BASED ON TEMPORAL PROFILES

In this Section, we present the smart card dataset that we use for our study and give the details of our approach to clustering public transportation passengers based on their temporal habits.

3.1 Case Study Dataset

We conduct our study on smart card data collected through the automated fare collection system of the Service des Transports en commun de l'Agglomération Rennaise (STAR). STAR operates a fleet of 65 bus lines and 1 subway line serving the metropolitan area of Rennes, which is the 11^{th} largest city in France with more than 200000 inhabitants. The operator established its automated fare collection system on March 1^{st}, 2006 and offers the possibility to travel on the network using a KorriGo smart card. As of 2011, 83% of the trips in STAR's transportation system were made using a KorriGo card.

The original dataset spans over the period from March 31^{st}, 2014 up to April 25^{th}, 2014, and contains both bus and subway boarding validations from more than 140000 KorriGo card holders. Each card is assigned a unique, anonymized identifier in order to ensure privacy and no personal information about the passengers were made available to us. Each validation contains, in addition to the card identifier, the date and time of the operation, the name and identifier of the boarding station, the name and identifier of the boarded bus or subway line, as well as information about the card type.

For our study, we only retain 53195 "regular" passengers. In order for a given smart card holder to qualify as a regular one, we check two conditions: the passenger (i) must have used his smart card for at least ten days during the studied period, and (ii) must have made his first boarding after 4 am each day at the same station 50% of the time (in which case the station is assigned as his "residential" station). The purpose of this step is to enhance the clustering results by removing occasional passengers for which the number of observed trips is insufficient and to have a rough estimate of the spatial location of the user residences. This is of interest since we also have access to a fine description of the socioeconomic characteristics of the city inhabitants through data released on a regular grid containing 200m per 200m cells by the French National Institute of Statistics and Economic Studies (INSEE).

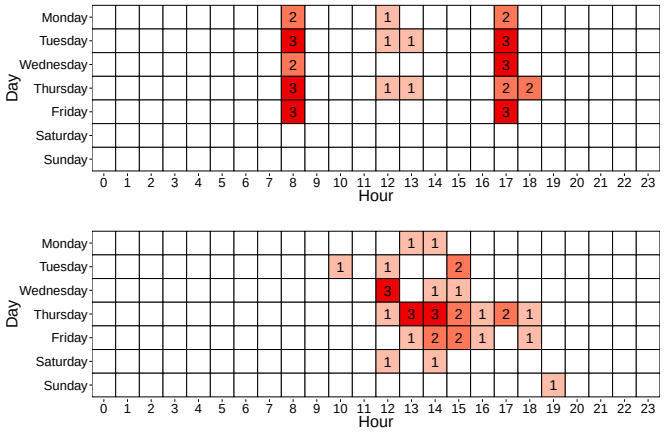

Figure 1: Temporal profiles of two passengers sampled from the smart card dataset.

3.2 Methodology

The aim of this part of our study is to discover groups of passengers who exhibit similar behaviors from a purely temporal standpoint (i.e. passengers taking public transportation at the same times without accounting for the boarding locations). Intuitively, the discovery of these groups can help identify frequent patterns in the way passengers use public transit and characterize the demand accordingly.

We start by aggregating each passenger's validations into a "weekly profile" describing the distribution of all his trips over each hour (0 through 23) of each day of the week (Monday through Sunday). Therefore, each passenger is an observation over 168 variables: the first variable is the number of trips he took on Monday 0 to 1 am, the second is the number of his trips on Monday 1 to 2 am, and so on. The temporal profiles of two passengers are illustrated in Figure 1 and we denote one such profile by $\mathbf{u}$.

Next, we cluster the passenger profiles. We do so by estimating a mixture of unigrams model [17] from our data, an approach often used to cluster documents in the context of information retrieval. Under this perspective, a passenger can be regarded as a "document" containing a collection of "words", with each word being, in our case, a combination of a day and an hour (e.g. Friday 10 am). In a (simple) unigram model, the words of each document are drawn independently from a single multinomial distribution. In a mixture of unigrams model, each document is first generated by selecting a cluster (called topic), then the words of that document are drawn from the conditional multinomial distribution relative to that specific cluster. Each cluster is thus described concisely as a distribution over the vocabulary. In our context, this translates to a cluster being the description of when trips

are more or less probable to be made. More formally, we use the following generative model for the data:

$$z \sim \mathcal{M}(1, \pi) \ , \quad (1)$$

$$\mathbf{u}|z \sim \mathcal{M}(D, \beta_z) \ . \quad (2)$$

with π the clusters' proportions, D the number of displacements recorded in $\mathbf{u}$ and β the cluster profiles. The parameters π and β were estimated from the data using a classical Expectation-Maximization (EM) algorithm. This maximizes the likelihood of the profiles which is derived from their distribution given by:

$$p(\mathbf{u}) = \sum_z p(z) \prod_{d=1}^{D} p(u_d|z) \ .$$

The approach was compared with possible alternatives that also lead to a compact representation of the passenger temporal profiles such as the Latent Dirichlet Allocation (LDA) [7], but eventually we chose to keep this simple approach for the ease of interpretation of the clustering results.

Another concern when estimating a mixture of unigrams model is that the number of clusters K must be specified in advance. In order to retrieve the model that best fits our data, we first use an EM algorithm to estimate the mixture of unigrams models while varying K from 2 to 30. To select an appropriate number of clusters, penalized likelihood criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used and asymptotically consistent but they are also known to be less efficient in practical situations than on simulated cases. To overcome this drawback in real situations, [6] have recently proposed a data-driven technique, called the "slope heuristic", to calibrate the penalty involving penalized criteria. The slope heuristic was first proposed in the context of Gaussian homoscedastic least squares regression and was since used in different situations, including model-based clustering. Birge et al. [6] proved the existence of a minimal penalty and that considering a penalty equal to twice this minimal penalty allows to approximate the oracle model in terms of risk. The minimal penalty is estimated in practice by the slope of the linear part of the objective function with regard to the model's complexity. A detailed overview and advices for implementation are given in [4].

3.3 Results

By proceeding as described earlier, we obtain a set of 16 clusters —that we refer to as "temporal clusters" from now on— each describing a temporal mobility pattern. In order to analyze the results, we can inspect visual representations of the daily temporal profiles obtained for each cluster (cf. Figure 2) and cross these visualizations with the proportions of fare types used by the passengers belonging to the corresponding clusters (cf. Figure 3).

As it can be seen on Figure 2, a first distinction can be made between two types of clusters: (i) clusters exhibiting routine behaviors (clusters 5 up to 16, except cluster 6), and (ii) clusters where the transport demand seems to be diffuse and does not portray typical home-to-work commutes during weekdays (clusters 1 up to 4). For example, one can clearly see that clusters 11, 12, and 14 are characterized by a morning peak of use at 7 am (resp. 8 am for cluster 14), an evening peak at 4 pm (resp. 5 pm for cluster 11), and a rush of usage at Wednesday 12 am (Wednesday is the day-off for

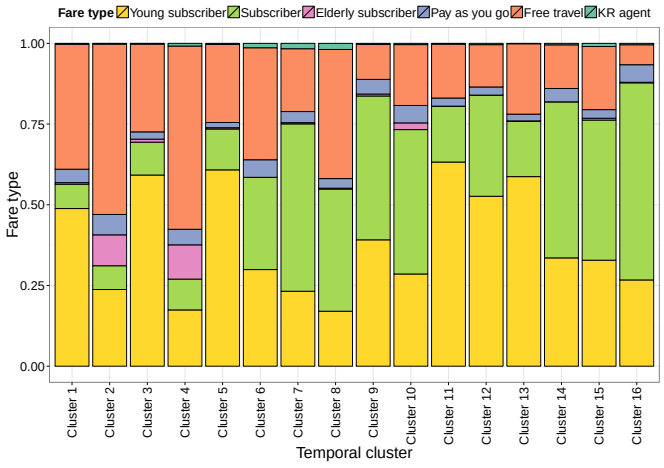

Figure 3: Proportions of fare types per temporal cluster.

students). Cluster 13 shows a quite similar behavior since it also highlights a morning peak at 7 am, but the usage on Wednesday and the evenings is more diffuse. We can also notice that for clusters 11, 12, and 14, no trips were made during the week-end, whereas cluster 13 exhibits a Saturday usage. Regarding the proportions of fare types per temporal profile of each cluster depicted in Figure 3, we can see that most of the passengers in clusters 11, 12, and 13 are made by " young subscribers", the difference in traveling behavior on Wednesday and evenings can be explained by the difference on daily commuting habits between school children and college or high school students for which leaving hours are more flexible. However, the proportion of "young subscribers" is less important in cluster 14 compared to the other clusters. This suggests that most of these passenger habits are indeed linked to daily school schedules but are rather made by parents aligning their trips with the school-leaving hours of their children.

Considering the remaining clusters (9, 10, 15 and 16) portraying daily routine behaviors, varying kinds of typical commuting patterns can be identified. Cluster 15 is related to passengers who travel in the morning peak (at 8 am), in the lunch break between 12 pm and 2 pm, and again in the evening rush hour between 5 pm and 6 pm. A quite similar behavior can be observed for passengers in cluster 16 which, in contrast, do not use transit during the lunch break. Regarding the proportions of fare types, " young subscribers" are largely represented in this cluster. Clusters 9 and 10 also exhibit a morning peak of usage at 9 am and evening peak at 6 pm and 7 pm (one-hour shift compared to Cluster 16). One of the main differences between these two clusters is the fact that passengers in cluster 10 use transit during Saturday more frequently than those in cluster 9.

Clusters 7 and 8 exhibit an early morning use (at 6 am for cluster 7, at 5 am for cluster 8) and no transit use during the week-end, which suggests that the trips are not made for leisure but are rather exclusively related to employment. The free travel fare type is higher for cluster 8. Cluster 5 exhibits daily routine travels. The peaks of transit use are early in the morning at 7 am, in the lunch break and in a

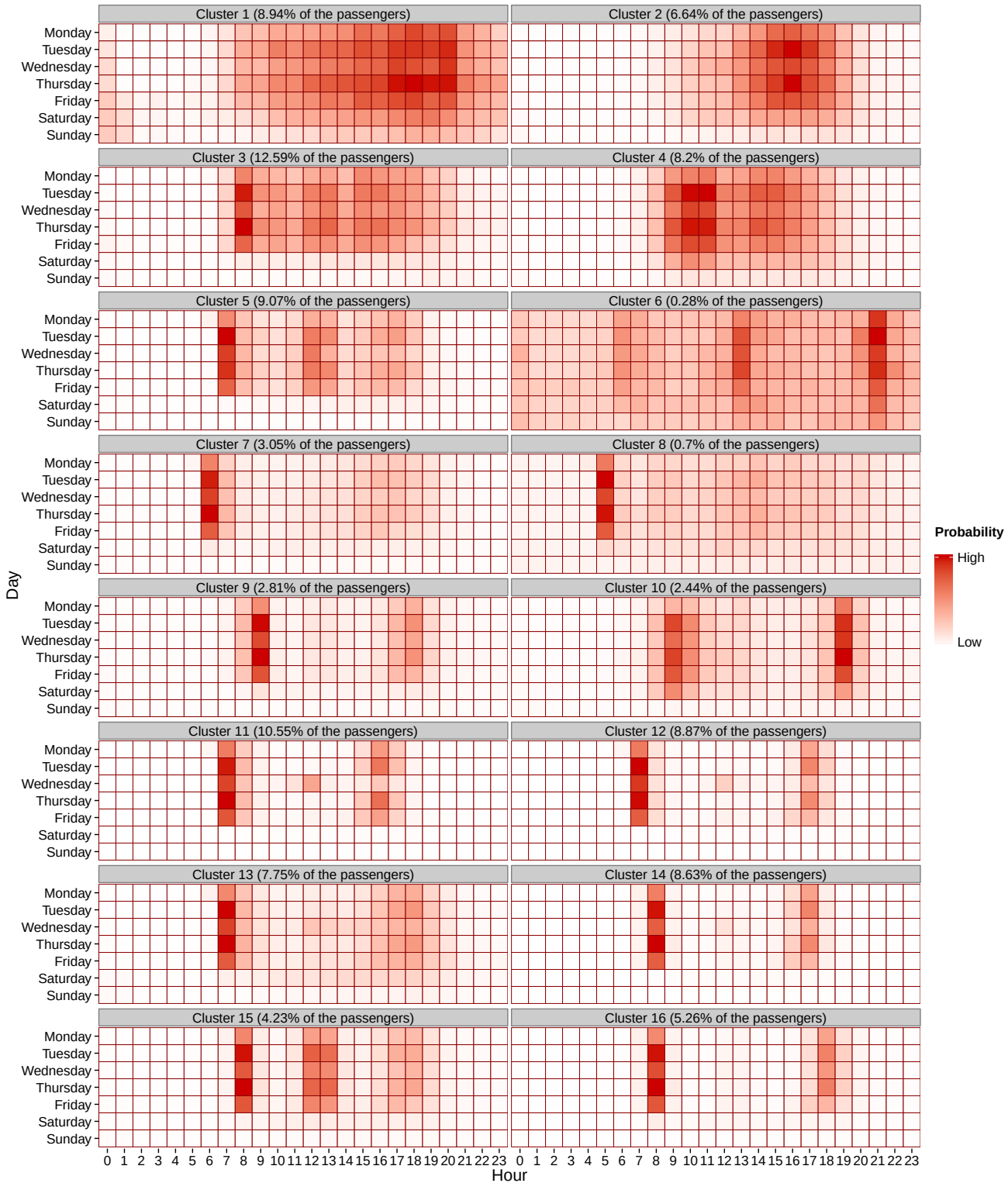

Figure 2: Temporal clusters estimated using the mixture of unigrams model.

diffuse manner in the evening. Passengers in this cluster do not use the transit in the week-end neither. Crossing these remarks with the fact that most of these users are "young subscribers", one can conclude that most of the passengers of this cluster are students.

Clusters 2 and 4 do not exhibit daily routines, meaning thus that such passengers travel after the morning rush hour and in a sporadic manner during the evenings. As it can be seen on Figure 3, these two clusters contain the highest proportions of elderly subscribers and passengers benefiting from free travel.

Finally, clusters 1 and 6 exhibit a diffuse daily usage of the transit system. However, they differ since some peaks of usage can be identified on the temporal cluster 6 such as 6 am, 1 pm, and 9 pm. Also, the proportions of fare types in this cluster are balanced between young subscribers, free travel and subscribers whereas cluster 1 mainly includes "free travel" and "young subscribers". The distinctive travel pattern in this cluster suggests that the passengers are laborers that work in rotating shifts.

4. ADDING CONTEXT THROUGH SOCIO-ECONOMIC DATA

We now discuss how socioeconomic data can be used to complement the study of temporal clusters.

4.1 Clustering of Socioeconomic Variables

In order to improve the analysis of the temporal clusters of passengers, we propose to introduce contextual information regarding the socioeconomic characteristics of the places of residence assigned to passengers through their "residential" stations. Socioeconomic data released on regular grids containing 200m per 200m cells by the French National Institute of Statistics and Economic Studies (INSEE) are available to conduct this study. The data contain several variables such as population count, income per consumption unit, cumulative living area of the households, etc.

Our intuition is to perform spatial clustering over these socioeconomic variables in order to obtain a synthetic description of the neighborhoods of the stations assigned to passengers. We address this spatial clustering problem using a Hidden Random Markov Field (HRMF) model that is well adapted to processing spatial data. The particularity of spatial data is that they cannot be viewed as a collection of independent variables because the values of the neighboring spatial entities are very often correlated one with another. HRMF models are often used for image segmentation or pixel-based classification [5] because they allow the incorporation of spatial dependencies in the building of a classifier through a neighboring system encoded by a Potts model for example. Spatial clustering in this context involves the estimation of the parameters of a hidden Markov field that corresponds to a discrete variable encoding the K clusters, and that must be discovered from the observation of the socioeconomic variables (supposedly following a multivariate Gaussian distribution). To estimate these parameters, several solutions exist based on the use of an EM-type algorithm. Notably, [9] proposed the NREM algorithm based on the mean-field approximation. It is designed for the parameters estimation of a HRMF in an unsupervised learning framework. Being able to fully estimate the parameters, including β the symmetric matrix of the Potts model encoding the spatial interactions between clusters, is an advantage that encourages us to use this particular algorithm because it gives more flexibility to recover the underlying clusters and since the clusters that we aim to identify may not have the same frequencies of apparition, they can have more or less grouped spatial structures. In the rest of this paper, the clusters retrieved as aforementioned will be referred to as "socioeconomic clusters".

Table 1: Mean and standard deviation of the socioeconomic variables in the socioeconomic clusters ($K = 7$).

Cluster	Area (m^2)	Income (€)	Population
Collective housing, Medium+ income	15696 (7763)	21889 (1791)	392 (194)
Collective housing, Low income	12510 (7553)	14499 (2364)	397 (244)
Individual housing, High density	8495 (2452)	20804 (2502)	239 (71)
Individual housing, High income	4359 (1847)	24422 (1899)	99 (45)
Individual housing, Medium inc. & dens.	3721 (1576)	20761 (2379)	103 (45)
Individual housing, Low density	695 (337)	21655 (2794)	17 (9)
Individual housing, Low density	167 (85)	21361 (2749)	4 (2)

N.B. standard deviation is indicated in parentheses.

We run the NREM algorithm on a set formed by the cells of the grid containing one or more stations and their neighboring cells using 8-D connectivity. This way we have a spatial coverage large enough to characterize the living neighborhood of the passengers, assuming here that passengers travel only a small distance from their home to the stations (less than 600 meters). We run the algorithm with a 7-colors Potts model. We choose here the number $K = 7$ manually according to the best interpretability of the clusters. Model selection based on log-likelihood criteria are difficult to apply here because we can only compute an approximated version of the log-likelihood and the existing approximated criteria [10] show inconsistent behaviors in our application. The mapping of the clustering results are visible on Figure 4. On this figure and for the crossing of the results with the temporal clustering, two clusters have been aggregate together because of their very similar profiles (cf. Table 1). The six remaining clusters provide a characterization of the residential neighborhoods in terms of level of income, population density, and type of buildings deduced from the combination of population and build area densities. We can observe from the mapping of the result that the positions of these clusters seem related to the distance to the center of the city of Rennes, which could have an influence on the analysis of the clusters describing the temporal behavior of the transportation operator's passengers.

4.2 Results

We start by examining the proportions of fare types used by the passengers of each of the socioeconomic clusters retrieved through clustering of socioeconomic variables (cf. Figure 5). The figure reveals some predictable results. For instance, we note that the highest proportion of passengers benefiting from free travel (which is attributed based on social and income considerations) is observed for the socioeconomic clus-

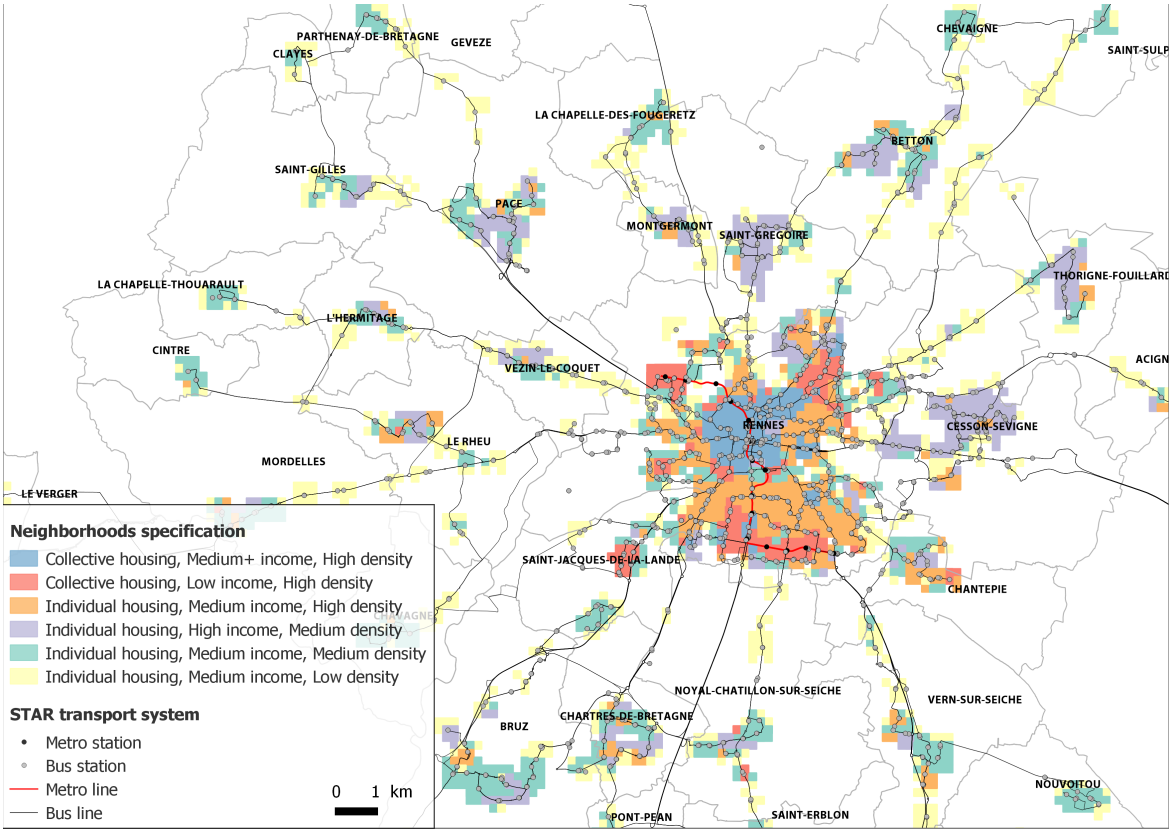

Figure 4: Socioeconomic clustering of the residential neighborhoods.

ter "Collective housing, low income, high density", whereas the lowest proportion of this fare type is in the "individual housing, high income medium density" cluster. The lowest proportion of young subscribers is also observed in the "Collective housing, low income, high density" cluster, which is also expected since —inter alia— students in this group logically use free travel instead of student-targeted fare types. The distribution of elderly subscribers seems balanced between clusters, except for the cluster of "Individual housing, Medium income, Low density" from which they are practically absent. This suggests that elderly people who used the public transport network live in medium to high density areas.

We also attempt to identify links between the socioeconomic clustering of the city and the temporal clustering performed on passengers profiles. To this effect, we inspect the proportions of socioeconomic clusters for each temporal cluster (cf. Figure 6). This figure shows that users belonging to cluster 8 which is characterized by an atypical early morning peak (5 am) are mainly located in "collective housing, low income, high density" areas, thus suggesting that this transport demand is made by early workers located in this kind collective housing (the corresponding spatial location can be seen on Figure 4). The same observation can be made for the temporal cluster 6 characterized by three peaks of usage at 6 am, 1 pm, and 9 pm. These peaks can correspond to laborers working nightshifts or daytime schedules involving three working teams.

Overall, the socioeconomic clusters seem to be similarly distributed across all temporal clusters. The highest demand

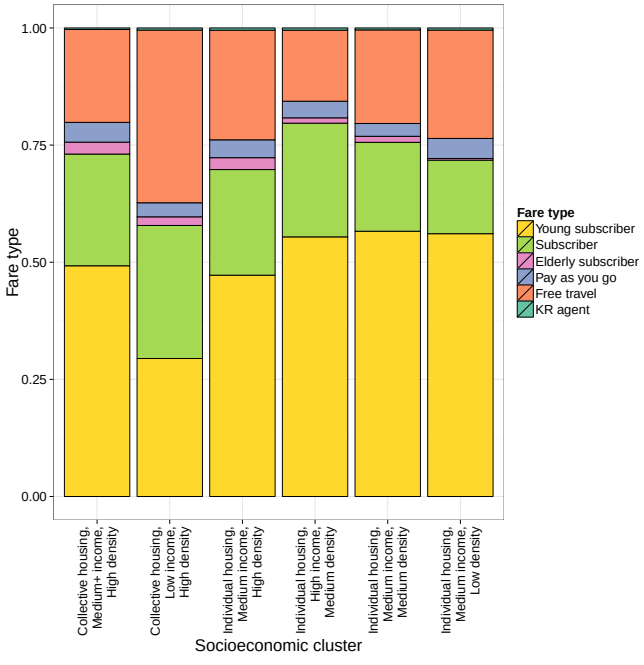

Figure 5: Proportions of fare types per socioeconomic cluster.

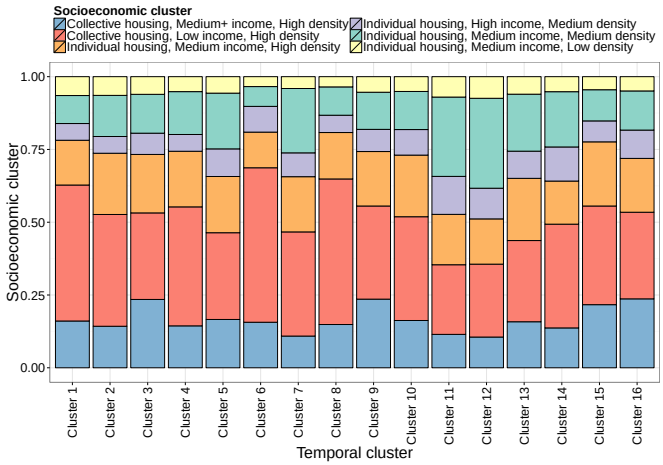

Figure 6: Proportions of socioeconomic clusters per temporal cluster.

for public transportation seems to come mainly from low income individuals and medium income individuals living in medium and high density areas. The lowest demands, on the other hand, are registered for high income individuals and medium income living in low density areas. Geographically, the latter two groups are located far from the center of Rennes, which suggests that they prefer using their private cars rather than using public transportation.

5. CONCLUSIONS

Smart card data present a unique opportunity to study passenger travel behavior in public transportation systems. In this paper, we introduced an approach to clustering passengers based on their temporal habits. By estimating a mixture of unigrams model from trips captured through smart card data, we retrieve weekly profiles (or temporal clusters) depicting different public transportation demands. By inspecting these profiles, it is possible to identify different groups of workers engaging in home-work commutes at different times of the day, students traveling to and from school, etc. We also made an attempt to link these profiles to socioeconomic data about the city in order to improve their interpretability. The results show, for example, which socioeconomic classes are more susceptible of using public transportation than the others, whether or not certain fare types are exclusively used by passengers located at particular areas of the city, etc.

Several improvements can be made based on the work presented herein. For example, the residential areas of the passengers are roughly estimated through the assignment to the most frequent first-departure station. Alternative, more robust approaches need to be investigated in order to enhance the accuracy of the pairing between smart card and socioeconomic data, and insure the quality of the interpretation of the results. We chose to cluster the passengers based solely on the boarding time of the trips they made. It would be interesting to include the spatial dimension (i.e. boarding locations) either during the clustering process or during the interpretation step (e.g. to identify the stations that are impacted by a given type of demand). The number of clusters discovered using our approach can be overwhelming to analyse. This issue can be addressed by trying to regroupe similar clusters and aggregating them into a hierarchy that is more suitable for multi-level exploration (i.e. start with a small number of coarse clusters to quickly understand the macro structure of passenger behavior, then expand interesting clusters to reveal more refined patterns). Also, we focused only on smart card data involving trips made by bus and subway. It would be interesting to study how these modes compare to and complement other transportation modes such as Bike Sharing Systems (BSS), etc. Finally, we adopted an exploratory, data-driven approach for this study. Consequently, our findings need to be examined by experts (urbanists and public transportation specialists) in order to confirm their adherence to reality.

6. ACKNOWLEDGMENTS

This work is undertaken as part of the Mobilletic project and is funded through the PREDIT (Programme de recherche et d'innovation dans les transports terrestres) program. The authors would like to thank Keolis Rennes who generously provided data for this study, especially Mr. Sébastien Leparoux for his active help and support.

7. REFERENCES

[1] B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. In *12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, 5 2006.

[2] M. Bagchi and P. R. White. What role for smart-card data from bus systems? *Proceedings of the ICE - Municipal Engineer*, 157:39–46(7), 2004.

[3] M. Bagchi and P. R. White. The potential of public transport smart card data. *Transport Policy*, 12(5):464–474, 9 2005.

[4] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[5] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[6] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[8] P. Bouman, E. van der Hurk, L. Kroon, T. Li, and P. Vervest. Detecting activity patterns from smart card data. In *25th Benelux Conference on Artificial Intelligence (BNAIC 2013)*, 2013.

[9] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

[10] F. Forbes and N. Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101, 2003.

[11] T. Fuse, K. Makimura, and T. Nakamura. Observation of travel behavior by ic card data and application to

transportation planning. In *Special Joint Symposium of ISPRS Commission IV and AutoCarto 2010*, 2010.

[12] N. Lathia and L. Capra. How smart is your smartcard? measuring travel behaviours, perceptions, and incentives. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 291–300, New York, NY, USA, 2011. ACM.

[13] N. Lathia, D. Quercia, and J. Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In *Proceedings of the 10th International Conference on Pervasive Computing*, Pervasive'12, pages 91–98, Berlin, Heidelberg, 2012. Springer-Verlag.

[14] N. Lathia, C. Smith, J. Froehlich, and L. Capra. Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, 9(5):643 – 664, 2013. Special issue on Pervasive Urban Applications.

[15] X.-l. Ma, Y.-J. Wu, Y.-h. Wang, F. Chen, and J.-f. Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36(0):1 – 12, 2013.

[16] C. Morency, M. Trépanier, and B. Agard. Analysing the variability of transit users behaviour with smart card data. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pages 44–49, 9 2006.

[17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134, May 2000.

[18] J. Y. Park, D.-J. Kim, and Y. Lim. Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea, 2008.

[19] M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

[20] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *In Workshop on Artificial Intelligence for Web Search (AAAI 2000*, pages 58–64. AAAI, 2000.

[21] W. Tran. Analysis of the differences in travel behaviour between pay as you go and season ticket holders using smart card data. In *1st Civil and Environmental Engineering Student Conference*, 6 2012.

[22] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology*, 2014.