CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-38

Simultaneous Dimension Reduction and Clustering via the NMF-EM Algorithm

L. CAREL[1]
P. ALQUIER[2]

[1] CREST-ENSAE, France. E-mail : Lena.carel@ensae.fr
[2] CREST-ENSAE, France. E-mail : Pierre.alquier@ensae.fr

Simultaneous Dimension Reduction and Clustering via the NMF-EM Algorithm

Léna CAREL[(1,2)], Pierre ALQUIER[(1)]

(1) CREST, ENSAE, Université Paris Saclay

5 avenue Henry Le Chatelier, 91120 Palaiseau - France

(2) Transdev Group

3 allée de Grenelle, 92442 Issy-les-Moulineaux CEDEX - France

Abstract

Mixture models are among the most popular tools for model based clustering. However, when the dimension and the number of clusters is large, the estimation as well as the interpretation of the clusters become challenging. We propose a reduced-dimension mixture model, where the K components parameters are combinations of words from a small dictionary - say H words with $H \ll K$. Including a Nonnegative Matrix Factorization (NMF) in the EM algorithm allows to simultaneously estimate the dictionary and the parameters of the mixture. We propose the acronym NMF-EM for this algorithm. This original approach is motivated by passengers clustering from ticketing data: we apply NMF-EM to ticketing data from two Transdev public transport networks. In this case, the words are easily interpreted as typical slots in a timetable.

Contents

1 Introduction **2**
- 1.1 Model-based clustering and urban computing 2
- 1.2 Dimension reduction in mixture models: variable selection 2
- 1.3 Dimension reduction in mixture models: the NMF-EM algorithm 3

2 The NMF-EM algorithm for mixture of multinomials **5**
- 2.1 The mixture of multinomials model 5
- 2.2 Explicit form of NMF-EM for mixture of multinomials 7
- 2.3 Discussion on the choice of H and K 8

3 Application to ticketing data **9**
- 3.1 Description of the data 9
- 3.2 Passengers profile clustering 10
- 3.3 Stations profile clustering 16
- 3.4 Passengers profile clustering on another network 22

4 Conclusion **25**

1 Introduction

1.1 Model-based clustering and urban computing

With the growing ability to collect and store data in transports system, electricity consumption and more, urban computing is becoming a major tool in urban policy and planning [37]. For example, for transports system, there is a growing litterature on ticketing and smart-card data processing in trains and buses [24, 26, 10, 28, 7], bike-sharing systems [30, 8, 5, 15] or taxis [27].

The range of machine learning and statistical tools used in urban computing is large. This goes from descriptive data-mining techniques as in [24] to statistical models. Model based clustering usually involve mixture models, that can be estimated by the EM algorithm. We refer the reader to [13] for an introduction to model based clustering, to [22] for an introduction to mixture models and to Chapter 9 in [3] for a general introduction to the EM algorithm (among others). In order to ease the following discussion, let us introduce our notations for mixture models. Assume that we are given a parametric family of distributions $(f_\theta)_{\theta\in\mathbb{R}^M}$. We assume the observations $Y_1,\dots,Y_n$ are i.i.d from a distribution of the form

$$\sum_{k=1}^{K} p_k f_{\theta_{\cdot,k}}(\cdot), \tag{1}$$

where each $\theta_{\cdot,k}\in\mathbb{R}^M$ is a column of a $K\times M$ matrix θ. Also, for the sake of concision, let $p=(p_1,\dots,p_K)\in\mathbb{R}^K$. A way to understand these models, that are useful for clustering purposes, is to introduce i.i.d hidden class variables: $Z_i=(Z_{i,1},\dots,Z_{i,K})\sim\mathcal{M}ult(p,1)$ (multinomial distribution). Then Y_i drawn from (1) can be obtained by:

$$Y_i\big|(Z_{i,k}=1)\sim f_{\theta_{\cdot,k}}(\cdot).$$

Mixture models were used by [10, 8, 5] for transports data, with very convincing results.

1.2 Dimension reduction in mixture models: variable selection

However, there are still a few issues with these models. When the dimension M is large, the estimation of the matrix θ, that is, of $M\times K$ parameters, is challenging from a computational perspective, moreover, the estimates are likely to have a large variance (curse of dimensionality). And more importantly in practice, it becomes more difficult to provide an interpretation to the cluster parameters $\theta_{\cdot,k}\in\mathbb{R}^M$. For example, [10] consider the Y_i's as passengers profile, that is, weekly timetables of transport usage by passengers - generated by multinomial distribution $f_{\theta_{\cdot,k}}(\cdot)$. Depending on the scale of the grid, the dimension of the profiles can be large. And it is indeed argued in [7] that some profiles in [10] are not easily interpretable. It is then necessary to reduce the dimension - that is, to seek for a matrix θ in a space with dimension much smaller than $M\times K$. Up to our knowledge, the only approach proposed to this purpose was variable

selection, that is, idendification of "useless" components in the Y_i's, leading to a reduction of the parameter space $\mathbb{R}^M$. This approach was used succesfully in many applications and is well-understood in theory [29, 32, 20, 21]; we refer the reader to the very nice and recent review [12] for a state-of-the-art and more references.

1.3 Dimension reduction in mixture models: the NMF-EM algorithm

Still, dimension reduction is not reducible to variable selection: we simply remind the reader that PCA performs indeed a reduction of the dimension without selecting variables. In the case where $\theta \in \mathbb{R}_+^{M\times K}$ we can also think of Nonnegative Matrix Factorization (NMF). Introduced by [17], NMF rewrites columns of a given matrix as positive combinations of elements in a small dictionary. These elements are often refered to as "words". It turns out that this dictionary is often easily interpretable. NMF was succesfully used in document clustering [34, 31], collaborative filtering and recommender systems on the Web [16], dictionary learning for images [17], topic extraction in texts [25] or time series recovering [23] - among others. NMF was also used as a tool in transports data in [15, 27, 28], but in a rather different way of what follows. In our setting, we would simply rewrite

$$\underbrace{\begin{pmatrix} \theta_{1,1} & \dots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \dots & \theta_{M,K} \end{pmatrix}}_{\theta} = \underbrace{\begin{pmatrix} \vartheta_{1,1} & \dots & \vartheta_{1,H} \\ \vdots & \ddots & \vdots \\ \vartheta_{M,1} & \dots & \vartheta_{M,H} \end{pmatrix}}_{\vartheta} \underbrace{\begin{pmatrix} \Lambda_{1,1} & \dots & \Lambda_{1,K} \\ \vdots & \ddots & \vdots \\ \Lambda_{H,1} & \dots & \Lambda_{H,K} \end{pmatrix}}_{\Lambda}$$

with $H \leq K$, under the assumption that all the entries in ϑ and Λ are nonnegative. When $H \ll K$, the dimension reduction is substantial.

In a previous work [7], we proposed to use an approximate NMF on the data $Y_1, \dots, Y_n$ and then to apply a (model-free) clustering algorithm on the decomposed observations. The improvement in terms of interpretability with respect to previous work was striking.

However, this approach was completely *ad hoc* and not satisfying from a theoretical perspective for two reasons. First, there are many possible criterion to approximate NMF: the Poisson-likekihood criterion used in [17, 18], the quadratic criterion (Gaussian-likelihood) used in Chapter 9 in [6] and in [18], the Ikuro-Saito divergence used in [11]... Secondly, as it was a model-free approach, it did not answer the question of the choice of the criterion. The choice of a model as (1) indeed imposes a natural criterion: namely, the likelihood. Our proposition is then simply to consider the mixture model with NMF. That is, $Y_1, \dots, Y_n$ are i.i.d from

$$g_{p,\vartheta,\Lambda}(\cdot) = \sum_{k=1}^{K} p_k f_{(\vartheta\Lambda)_{\cdot,k}}(\cdot) \tag{2}$$

parametrized by p, Λ and ϑ. Note that we can still introduce i.i.d random variables $Z_i = (Z_{i,1}, \dots, Z_{i,K}) \sim \mathcal{M}ult(p, 1)$ (multinomial distribution). Then Y_i drawn from (2) can be obtained by:

$$Y_i \big| (Z_{i,k} = 1) \sim f_{(\vartheta\Lambda)_{\cdot,k}}(\cdot).$$

For short, put $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$. The log-likelihood is given by

$$\ell(\vartheta, \Lambda, p|Y) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p_k f_{(\vartheta\Lambda)_{\cdot,k}}(Y_i) \right).$$

Note that, in some sense, this could be seen as a bridge between "model free clustering" relying on NMF or spectral clustering as in [9, 36] and model-based clustering as in [13]. Of course, the maximization of this log-likelihood is not easier than for (usual) mixture models, that are actually a special case of (2) (that can be seen by taking $H = K$ and $\Lambda = I_K$, the identity matrix of size K). We then propose to adapt the EM algorithm to this setting. We introduce the completed log-likelihood

$$\ell(\vartheta, \Lambda, p|Y, Z) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{i,k} \log \left(p_k f_{(\vartheta\Lambda)_{\cdot,k}}(Y_i) \right).$$

A step of the EM algorithm, given current parameters $(\vartheta^{(c)}, \Lambda^{(c)}, p^{(c)})$ would then be as follows.

E-step Put

$$\begin{aligned} Q^{(c)}(\vartheta, \Lambda, p) &= \mathbb{E}_{\vartheta^{(c)},\Lambda^{(c)},p^{(c)}}[\ell(\vartheta, \Lambda, p|Y, Z)|Y] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\vartheta^{(c)},\Lambda^{(c)},p^{(c)}}[Z_{i,k}|Y] \log \left(p_k f_{(\vartheta\Lambda)_{\cdot,k}}(Y_i) \right) \end{aligned}$$

and note that

$$t_{i,k}^{(c)} := \mathbb{E}_{\vartheta^{(c)},\Lambda^{(c)},p^{(c)}}[Z_{i,k}|Y] = \frac{p_k^{(c)} f_{(\vartheta^{(c)}\Lambda^{(c)})_{\cdot,k}}(Y_i)}{\sum_{k'=1}^{K} p_{k'}^{(c)} f_{(\vartheta^{(c)}\Lambda^{(c)})_{\cdot,k'}}(Y_i)}. \tag{3}$$

M-step Compute

$$(\vartheta^{(c+1)}, \Lambda^{(c+1)}, p^{(c+1)}) := \underset{\vartheta_{j,h},\Lambda_{h,k} \geq 0}{\arg\max}\ Q^{(c)}(\vartheta, \Lambda, p). \tag{4}$$

Obviously, the challenging step is the M-step. While we obviously have, for $k \in \{1, \ldots, K\}$,

$$p_k^{(c+1)} = \frac{\sum_{i=1}^{n} t_{i,k}^{(c)}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} t_{i,k'}^{(c)}},$$

the nonnegativity constraint on ϑ and Λ makes the optimization with respect to these two matrices much harder. Many options might be possible, depending on the form of the density functions $f_u(\cdot)$. In general we suggest to use an alternating optimization method with respect to ϑ and Λ as done in NMF [17, 18]. Since then, many algorithms were proposed for NMF, and most rely on alternate optimization. We cannot list them all as the number of variants is huge. We mention ADMM described in [6, 33], Bayesian versions studied in [25,

1], using variational approximations and Monte-Carlo methods respectively... We also refer the reader to [19] for a numerical comparison of the different methods. In many applications, the simplest algorithm - the multiplicative method of [17, 18] - is very efficient. So this is the method we will use in Sections 2 and 3. This method iterates a step in ϑ, and a step in Λ. Each step is shown to improve the fit criterion in [18]. Note that the author claims that it also leads to convergence, but as argued in [14] the proof of this fact is actually incomplete. Some theoretical work is still needed there.

By now we already have introduced our algorithm and we hope that the reader has a clear idea of its general motivations. The end of the papers is organized as follows. We remind that our objective was the analysis of ticketing data. In Section 2 we explicit the model, and the NMF-EM algorithm, in the case of mixture of multinomials meant to represent temporal passengers profiles as in [10] - note that we will use this to represent stations profiles as well. We then present results on ticketing data from two networks provided by the Transdev Group in Section 3.

2 The NMF-EM algorithm for mixture of multinomials

2.1 The mixture of multinomials model

In [10] the authors propose to model a passenger temporal profile by a mixture of multinomial distribution. Namely, the time and days of smart card validations of a passenger i are recorded over a period of time (e.g. 1 month). The numbers of travels, N_i, is not our variable of interest, and will be considered as deterministic. We obtain as a result a vector

$$X_i = (X_{i,1}, \dots, X_{i,M})^T \in \mathbb{R}^M$$

where each coordinates represents the number of travels at a given pair time-day during the considered period. We consider a hourly grid, that is, Mon-12am, Mon-1am, etc... to Sun-11pm, with means that $M = 7 \times 24 = 168$. An example of traveler profile is given in Figure 1.

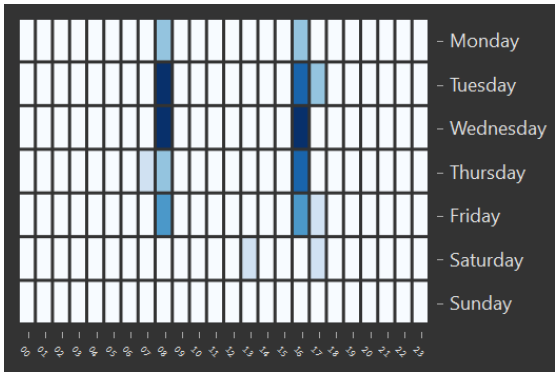


Figure 1: Temporal profile of a network user, taken from the data described in Section 3. Opacity is proportional to the number of smart-card validations.

It is natural to assume that there are clusters of passengers with rather similar profiles: for examples, employees in the same company or students in the same class are likely to commute at similar times. We introduce hidden variables for the clusters, $Z_i \in \{0,1\}^K$ with $Z_{i,k} = 1$ when individual i belongs to cluster k. Assuming that the average profile of cluster k is given by $\theta_{\cdot,k} \in \mathbb{R}^M$ we write

$$X_i \big| (Z_{i,k} = 1) \sim \mathcal{M}ult(\theta_{\cdot,k}, N_i).$$

However, according to our factorization assumption: $\theta = (\theta_{\cdot,1}| \dots |\theta_{\cdot,K}) = \vartheta\Lambda$ where ϑ is $M \times H$ and Λ is $H \times K$ for some $H \leq K$, we can rewrite:

$$X_i \big| (Z_{i,k} = 1) \sim \mathcal{M}ult((\vartheta\Lambda)_{\cdot,k}, N_i).$$

And the model is:

$$g_{p,\vartheta,\Lambda}(X_i) = \sum_{k=1}^{K} p_k \left[N_i! \prod_{j=1}^{M} \frac{(\vartheta\Lambda)_{j,k}^{X_{i,j}}}{X_{i,j}!} \right].$$

The log-likelihood is given by

$$\ell(\vartheta, \Lambda, p|X) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} p_k \left[N_i! \prod_{j=1}^{M} \frac{(\vartheta\Lambda)_{j,k}^{X_{i,j}}}{X_{i,j}!} \right] \right\}.$$

Note that the $\theta_{\cdot,k} = (\vartheta\Lambda)_{\cdot,k}$ are in the simplex

$$\mathcal{S}_M = \{(t_1, \dots, t_M) \in \mathbb{R}_+^M : t_1 + \dots + t_M = 1\}.$$

So we impose the same restriction on $\vartheta_{\cdot,k}$ and on the $\Lambda_{\cdot,j}$. Let $\mathcal{M}_{M,H,K}$ denote the set of all pairs (ϑ, Λ) of matrices $M \times H$ and $H \times K$ respectively, with $\vartheta_{\cdot,k}, \Lambda_{\cdot,j} \in \mathcal{S}_M$ for any k and j. Note that we actually have $H(M-1) + K(H-1) + K - 1$ degrees of freedom for $(\vartheta, \Lambda) \in \mathcal{M}_{M,H,K}$ and $p \in \mathcal{S}_K$, a fact that is useful to compute model selection criterion like AIC and BIC (see the discussion on model selection below).

Remark 2.1. *The reader might notice some similarity with the Latent Dirichlet Allocation (LDA) model in [4]. Indeed, there is a similar "two layers of mixtures" structure, in the sense that the parameters $(\theta_{\cdot,k})_{1\leq k\leq k}$ are themselves mixtures of a small number H of hyperparameters $(\vartheta_{\cdot,h})_{1\leq h\leq H}$ (also refered to as "words"). Then, the observations are drawn from distributions depending on the $(\theta_{\cdot,k})_{1\leq k\leq k}$. The main difference with [4] is the mixture structure in the distribution of the X_i's: we indeed assume that many passengers will actually have similar temporal profiles. In [4] there is no assumption that some groups of documents could have the same topic distribution.*

Let us denote $(\hat{\vartheta}, \hat{\Lambda}, \hat{p})$ the MLE, that is, a maximizer, for $(\vartheta, \Lambda) \in \mathcal{M}_{M,H,K}$ and $p \in \mathcal{S}_K$, of $\ell(\vartheta, \Lambda, p|X)$. This is our ideal estimator. The aim of the NMF-EM algorithm is to approximate $(\hat{\vartheta}, \hat{\Lambda}, \hat{p})$.

2.2 Explicit form of NMF-EM for mixture of multinomials

From (3), values $t_{i,k}^{(c)}$ are given by

$$\begin{aligned} t_{i,k}^{(c)} &= \frac{p_k^{(c)} N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \vartheta_{j,h}^{(c)} \Lambda_{h,k}^{(c)}\right)^{X_{i,j}}}{X_{i,j}!}}{\sum_{k'=1}^{K} p_{k'}^{(c)} N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \vartheta_{j,h}^{(c)} \Lambda_{h,k'}^{(c)}\right)^{X_{i,j}}}{X_{i,j}!}} \\ &= \frac{p_k^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \vartheta_{j,h}^{(c)} \Lambda_{h,k}^{(c)}\right)^{X_{i,j}}}{\sum_{k'=1}^{K} p_{k'}^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \vartheta_{j,h}^{(c)} \Lambda_{h,k'}^{(c)}\right)^{X_{i,j}}}. \end{aligned}$$

We have

$$\begin{aligned} Q^{(c)}(\vartheta, \Lambda, p) &= \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k}^{(c)} \log \left(p_k N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k}\right)^{X_{i,j}}}{X_{i,j}!} \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k}^{(c)} \Bigg[\log(p_k) + \log(N_i!) \\ &\qquad + \sum_{j=1}^{M} \left(X_{i,j} \log \left(\sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k} \right) - \log(X_{i,j}!) \right) \Bigg]. \end{aligned}$$

As mentioned earlier,

$$p_k^{(c+1)} = \frac{\sum_{i=1}^{n} t_{i,k}^{(c)}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} t_{i,k'}^{(c)}},$$

and

$$(\vartheta^{(c+1)}, \Lambda^{(c+1)}) = \underset{(\vartheta,\Lambda) \in \mathcal{M}_{M,H,K}}{\arg\max} \sum_{k=1}^{K} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} X_{i,j} t_{i,k}^{(c)} \right) \log \left(\sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k} \right).$$

For ease of reading, let us note $M_{j,k}^{(c)} = \sum_{i=1}^{n} X_{i,j} t_{i,k}^{(c)}$ for short. Then the previous equation becomes

$$(\vartheta^{(c+1)}, \Lambda^{(c+1)}) = \underset{(\vartheta,\Lambda) \in \mathcal{M}_{M,H,K}}{\arg\max} \sum_{k=1}^{K} \sum_{j=1}^{M} M_{j,k}^{(c)} \log \left(\sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k} \right). \tag{5}$$

To minimize this criterion under the constraint is not straightforward. Note that the maximization in (5) is equivalent to the minimization of

$$D(M^{(c)} || \vartheta\Lambda) := -\sum_{k=1}^{K} \sum_{j=1}^{M} \left\{ M_{j,k}^{(c)} \log \left(\sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k} \right) - \sum_{h=1}^{H} \vartheta_{j,h} \Lambda_{h,k} \right\}. \tag{6}$$

Indeed, for $(\vartheta, \Lambda) \in \mathcal{M}_{M,H,K}$ we have

$$\sum_{k=1}^{K}\sum_{j=1}^{M}\sum_{h=1}^{H}\vartheta_{j,h}\Lambda_{h,k} = \sum_{j=1}^{M}\sum_{k=1}^{K}(\vartheta\Lambda)_{j,k} = \sum_{j=1}^{M}1 = M$$

that does not depend on (ϑ, Λ). This is time to consider the popular NMF algorithms mentioned in the introduction: indeed, the multiplicative method proposed in [18] exactly aims at minimizing the divergence $D(M^{(c)}||\vartheta\Lambda)$ with respect to matrices ϑ and Λ with nonnegative entries (without the simplex constraint). We thus propose to iterate alternatively one multiplicative update of [18] followed by a proper renormalization of the matrices ϑ and Λ. We end up with the NMF-EM algorithm for mixture of multinomials summarized in Algorithm 1 page 8.

Algorithm 1 NMF-EM

Fix $\epsilon > 0$. Choose arbitrary $\vartheta^{(0)}$, $\Lambda^{(0)}$ and $p^{(0)}$; $c := 0$, CRIT $:= \infty$.
while $|\ell(\vartheta^{(c)}, \Lambda^{(c)}, p^{(c)}) - \text{CRIT}| > \epsilon$ **do**
  CRIT $:= \ell(\vartheta^{(c)}, \Lambda^{(c)}, p^{(c)})$.
  For all $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$,

$$t_{i,k}^{(c)} := \frac{p_k^{(c)} \prod_{j=1}^{M}\left(\sum_{h=1}^{H}\vartheta_{j,h}^{(c)}\Lambda_{h,k}^{(c)}\right)^{X_{i,j}}}{\sum_{k'=1}^{K} p_{k'}^{(c)} \prod_{j=1}^{M}\left(\sum_{h=1}^{H}\vartheta_{j,h}^{(c)}\Lambda_{h,k'}^{(c)}\right)^{X_{i,j}}} \text{ and } p_k^{(c+1)} =: \frac{\sum_{i=1}^{n} t_{i,k}^{(c)}}{\sum_{i=1}^{n}\sum_{k'=1}^{K} t_{i,k'}^{(c)}}.$$

  $\forall j,k \quad M_{j,k}^{(c)} = \sum_{i=1}^{n} X_{i,j} t_{i,k}^{(c)}$.
  Initialization of ϑ and Λ (arbitrarily), $q := \infty$.
  while $|Q^{(c)}(\vartheta, \Lambda, p^{(c+1)}) - q| > \epsilon$ **do**
    $q := Q^{(c)}(\vartheta, \Lambda, p^{(c+1)})$.
    $\forall h,k \quad \Lambda_{h,k} \leftarrow \Lambda_{h,k} \frac{\sum_j \vartheta_{j,h} M_{j,k}^{(c)}/(\vartheta\Lambda)_{j,k}}{\sum_j \vartheta_{j,h}}$
    $\forall h,k \quad \Lambda_{h,k} \leftarrow \frac{\Lambda_{h,k}}{\sum_{k'} \Lambda_{h,k'}}$
    $\forall j,h \quad \vartheta_{j,h} \leftarrow \vartheta_{j,h} \frac{\sum_k \Lambda_{h,k} M_{j,k}^{(c)}/(\vartheta\Lambda)_{j,k}}{\sum_k \Lambda_{h,k}}$
    $\forall j,h \quad \vartheta_{j,h} \leftarrow \frac{\vartheta_{j,h}}{\sum_{h'} \vartheta_{j,h'}}$
  end while
  $(\vartheta^{(c+1)}, \Lambda^{(c+1)}) := (\vartheta, \Lambda)$.
  $c := c + 1$.
end while

2.3 Discussion on the choice of H and K

The choice of K is not a straightforward issue in mixture models. *A fortiori* the choice of the pair (H, K) is not easier.

First, notice that we computed the degrees of freedom of our model, that is

$H(M-1)+K(H-1)+K-1$. Thus, we can derive the AIC criterion

$$\text{AIC} = \ell(\hat{\vartheta}, \hat{\Lambda}, \hat{p}|x) - \frac{H(M-1)+K(H-1)+K-1}{2}$$

or the BIC criterion

$$\text{BIC} = \ell(\hat{\vartheta}, \hat{\Lambda}, \hat{p}|x) - \frac{[H(M-1)+K(H-1)+K-1]\log(n)}{2}.$$

The BIC criterion is widely used to choose the number of components in mixtures, we refer the reader to the aforementionned paper [5] for an example of application to transports data. However, note that conditions for the consistency of AIC and BIC is understood in some models, but definitely not in mixture models. These criterion are *ad hoc* in this case. Criterion more suitable for mixtures were investigated, like NEC and variants [2]. The slope heuristic used in [10] for mixture of multinomial gives nice results. More importantly, one has to remind what is the objective of the statistician. One usually think of AIC as a criterion to balance optimally bias and variance while BIC is supposed to identify the true model, when there is one [35]. This last assumption is obviously wrong - note that the multinomial model assumes that the different travels of a given passenger are independent, which is obviously an approximation! We believe that in many applications, interpretability of the results might be the criterion of reference.

3 Application to ticketing data

3.1 Description of the data

The data used in our study are the validations made during the month of September 2015 on one Transdev network in a medium size city. Ticketing data are the information obtained at each transaction made by a smart card on a validator system. For privacy reasons it is not possible to connect each validation to the user that made it. The feature that allows us to realize our study and create temporal profiles is a card number which is encrypted, and re-initialized every three months. It is thus impossible to follow the long-term behaviour of a user. This is the reason why we focus on a one month period. This period (September) have been chosen because it has no vacation nor bank holiday. MoÃğreover, we applied the method on data from another period (2015, January 14^{th} to February 28^{th}) and results were similar. During September 2015, more than $4,000,000$ check-ins have been made on the network by $232,430$ passengers. We call check-in every transaction made by an user with his smart card. Indeed, if a traveler make a trip with with a connection, two check-ins would be counted.

The data are agregated so that for each traveler, for each day of the week (Monday to Sunday) and each hour (00 to 23), we have the number of validation during the studied period. A passenger profile is thus defined by $24 * 7 = 168$ features. Figure 1 page 5 already provided an example of a temporal profile of one of the users. This traveler uses mainly the network at 8 a.m and 4 p.m.

We used the same strategy to create stations profiles: for each station, for each day of the week and each hour of the day, we know the number of validations

that occured at this station during the study period. In Figure 2, we show the temporal profile of the courthouse station, a tramway station in the city center. This station has travelers all day long, but knows an attendance peak every day from 4 to 6 p.m.

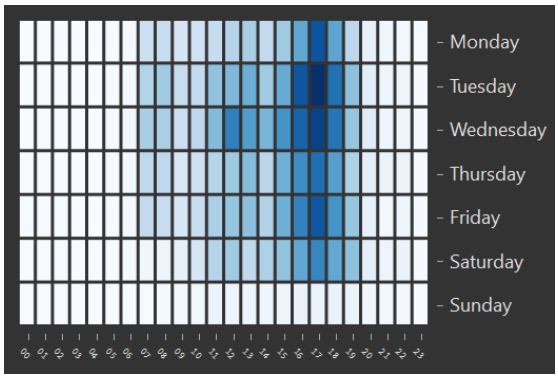


Figure 2: Temporal profile of the courthouse station

In order to avoid occasional users, that would not use enough their smart card to exhibit a clear pattern, data have been cleaned. We define a "regular card holder" as a card holder who

- travelled at least four days during September 2015 (so in particular we have $N_i \geq 4$);
- made his first boarding after 4 a.m each day at the same station 50% of the time.

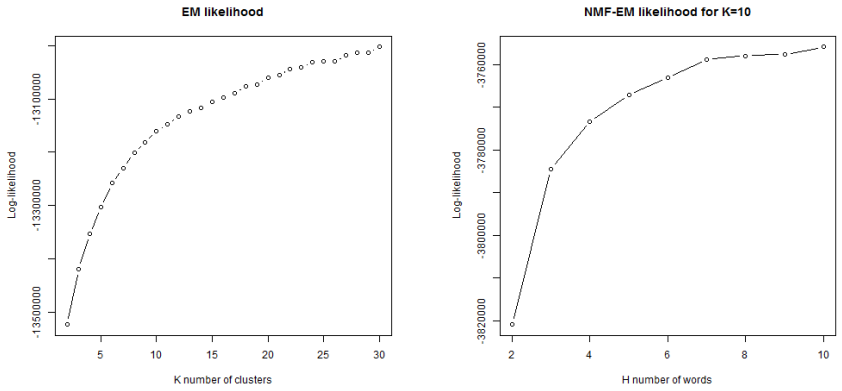
We only kept regular card holders for our analysis. After this cleaning step, we end up with $72,359$ profiles of passengers, which represent a bit more than $3,000,000$ check-ins – that means 31% of passengers represent 75% of check-ins. We also have 475 stations profiles.

3.2 Passengers profile clustering

We first focus on passengers profiles clustering. This allows us to create groups of people that have similar temporal habits. The method used to create these clusters is the NMF-EM algorithm from Section 2.

To choose the parameters H and K, we begin with the analysis of the log-likelihood of our model when $H = K$ for $K = 2 \dots 30$. Note that the estimation of the model in this case can be made by the usual EM algorithm for multinomial mixture model. Figure 1a shows the evolution of the log-likelihood as a function of K. This function clearly exhibits a linear behavior when $K \geq 10$. Thus, the slope heuristic suggests to consider $K = 10$.

Keeping now $K = 10$ fixed, we chose the value of H in the same way. First, we plot the log-likelihood as a function of H in Figure 1b.Then, by using again the slope heuristic method, we choose $H = 5$.



(a) Log-likelihood as a function of K, under $H = K$

(b) Log-likelihood as a function of $H \in \{2, ..., K\}$ under $K = 10$

Table 1: Model selection for the french network

The $H = 5$ words and the $K = 10$ clusters are represented in Figure 3 and in Figure 5 respectively. Remind that each cluster can be decomposed as a convex combination of words, some of them might have a null weight. For example, Figure 4 shows how the parameter of Cluster 5 can be written as a convex combination of words 4 and 2.

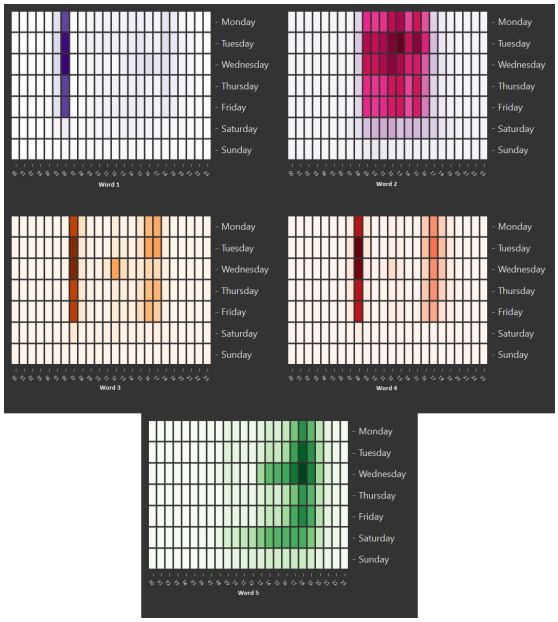


Figure 3: Words obtained by NMF-EM on users data with $K = 10$ and $H = 5$.

The interpretation of the words is direct:

1. Word 1: travels between 6 a.m and 7 a.m.

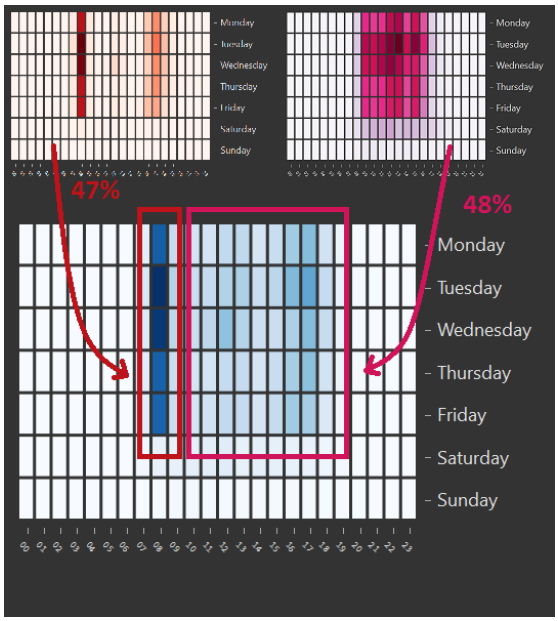


Figure 4: Decomposition of cluster 5 from words 4 and 2.

2. Word 2: diffuse component during off-peak periods (i.e. from 9 a.m to 4 p.m).
3. Word 3: travels at school hours. Indeed it is composed of travels between 7 and 8 a.m and between 4 and 5 p.m, except on Wednesdays, when the afternoon travel is replaced by one at noon.
4. Word 4: travels between 8 and 9 a.m.
5. Word 5: late afternoon peak, from 5 to 7 p.m, and Wednesdays and Saturdays afternoon.

We now attempt an interpretation of the clusters:

1. Clusters 1, 3, 4 and 6 present high travel probabilities in the morning and in the afternoon except Wednesdays where the afternoon travel is replaced by a higher probability of travel around noon. These four clusters are typical of French schools and high-schools hours. The main differences are:
   (a) Cluster 1: travels at 7 a.m and around 4 or 5 p.m.
   (b) Cluster 3: travel a bit more at 8 a.m.
   (c) Cluster 4: travelers are less susceptible to travel after 5 p.m.
   (d) Cluster 6: travels at 6 and 7 a.m.
2. Cluster 5: travels at 8 a.m and at 4 or 5 p.m.
3. Cluster 7: travels mainly at 6 a.m.
4. Cluster 9: travels at 8 a.m and at 5 p.m.

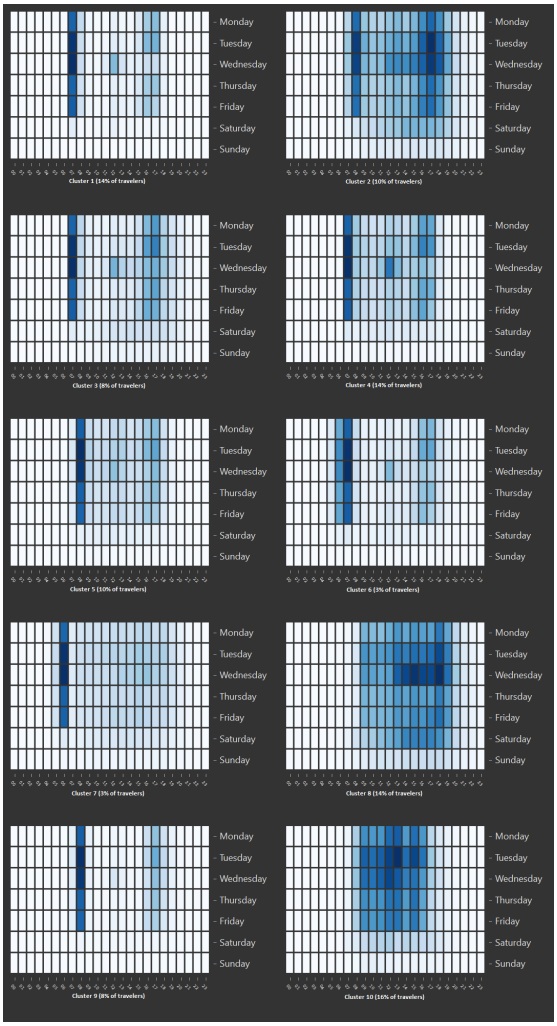


Figure 5: Clusters obtained by NMF-EM on users data with $K = 10$ and $H = 5$.

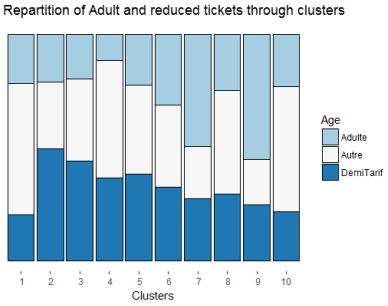
5. Clusters 2, 8 and 10: diffuse travel habits.

   (a) Cluster 2: travels Mondays to Saturdays from 7 a.m to 7 p.m with highest probabilities at 8 a.m and 5 p.m Mondays to Fridays.

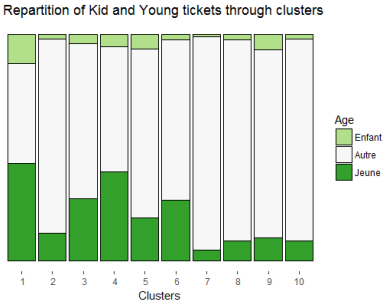
   (b) Cluster 8: diffuse travels Mondays to Saturdays from 9 a.m to 7 p.m.

   (c) Cluster 10: travels Mondays to Fridays from 9 a.m to 4 p.m.

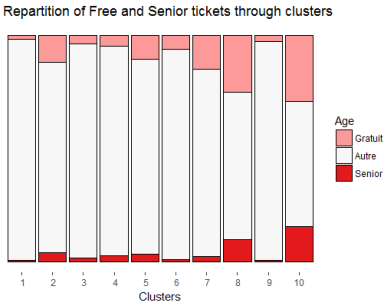
In order to explain the small differences that exist between some cluster, we need more informations about the travelers in them. As written above, we have no personal information in our data. Therefore, we are not able to describe individually the users in each cluster. However, for each transaction made, we have the encrypted card number and the transport ticket used. So we can recover for each card the most used transport ticket during the period. This provides interesting information as some schemes are associated to age ranges (Young, Senior...) and to time periods (Unit, Annual or Monthly Subscription). Let us now provide the description of each cluster in terms of age ranges (Figures 2a to 2c in Table 2). Note that in each figure, only two age ranges are represented and the other four are aggregated in an "Autre" category.



(a) Adults ("Adulte") and Reduced tickets ("DemiTarif")



(b) Kids ("Enfant") and Youngs ("Jeune")



(c) Free travelers ("Gratuit") and Elderly ("Senior")

Table 2: Age range analysis of the clusters

Adults are more present in clusters 7 and 9, that are clusters with check-ins mostly in the morning. People benefiting from half-price are present in every cluster but with highest rates in clusters 2, 3, 4 and 5. Children (4 to 6) are not very present on the network, but they are more represented in clusters

1, 5 and 9. Young travelers (6 to 25) are more present in clusters 1 and 4. These clusters correspond to scholar time slot. In clusters 8 and 10 there are large rate of seniors and free travelers. As these clusters have profiles of diffuse travels during the week and as free travelers are unemployed or low salaries people, these regroupments make sense.

Figure 6 shows the repartition of transport ticket type through clusters. Unit

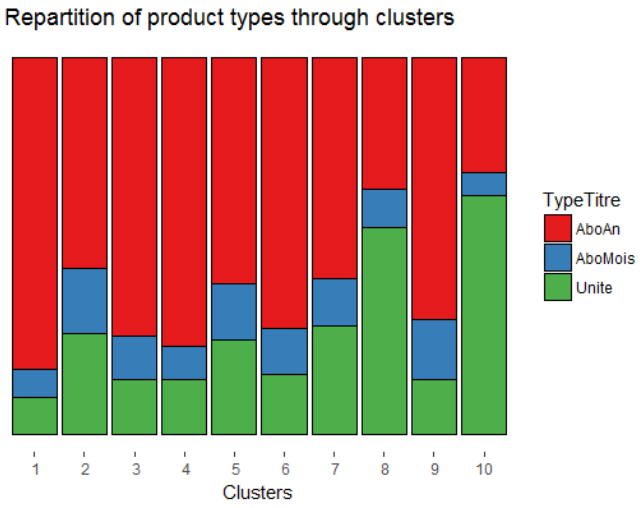


Figure 6: Transportation ticket type analysis of the clusters.

products are more used in clusters 8 and 10 that are clusters with a lots of seniors and free travelers. As they don't have obligations, they likely use unit products for occasional trips. Clusters 1, 3, 4 and 9, that have mostly scholar profiles although have a large majority of annual subscripters. A possible interpretation is that schoolchildren and students are public transportation captives, and have to use the network in order to go to class every day. Thus, buying an annual pass is more advantageous than buying any other product type.

As described in Subsection 3.1, we kept only users whose first trip of the day is made at the same station at least 50% of the study time. That main "morning station" is thus called the "home station" as it gives us an estimation of the residence place of users. In Tables 8 and 9 in Appendix A, we can observe the shares of clusters by home stations. It shows the share of travelers identified as belonging to every cluster living near each station.

We note that:

1. Cluster 1: travelers are over represented at peripheral stations.
2. Cluster 2: no particular pattern observed.
3. Cluster 3: no particular pattern observed.
4. Cluster 4: few stations show over representation of cluster 4.
5. Cluster 5: over representation of the cluster at two stations in the north.
6. Cluster 6: no particular pattern observed.

7. Cluster 7: One station is 100% represented by cluster 7. As only one user is assigned to this station, no particular pattern is observed.

8. Cluster 8: the cluster is over represented at one station in the city center and at another further.

9. Cluster 9: cluster 9 is over represented in few stations in the center.

10. Cluster 10: cluster is over represented in poorest neighborhoods of the city.

As the previous results show no geographical discrimination for some clusters, we add some contextual data to the stations. Thanks to the French National Institute of Statistics and Economis Studies (INSEE), there are open data permitting us to introduce more information. Firstly, a database containing socioeconomic data on a grid of $200m \times 200m$ is available. We extracted two indicator of it: the density of population and the percentage of households living in collective housing per tiles. Secondly, we used a database referencing and geolocating every french company or administration, named "SIRENE base". At last, we used two databases geolocating middle and high schools and universities and referencing the number of students in each educational establishment. These contextual data are represented in Table 3.

We can observe in Figure 3a that density of population is high in the city center, and decreases the further away we are. In Figure 3b, that there are more households in collective housing in the city centers and in few neighborhood. In Figure 3c, we observe that it is in the city center that incomes are the highest. Even if incomes decreases away from the city center, like the population density, highest incomes are more condensed. As for the other indicators, jobs are most concentrated in the city center (Figure 3d).

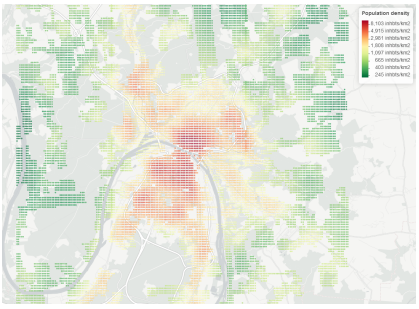
In order to improve the description of the clusters, we created catchment areas around the stations according to the type of vehicle (subway, tramway, bus).Then, we were able to get an average profile of the inhabitants living in each catchment area. Thus, it becomes possible to obtain a description of the clusters.

In Table 4, we observe on Figure 4a that travelers from clusters 1 and 6 are coming from neighborhoods with lower density of population. On Figure 4b, we notice that travelers from clusters 1 and 6 are more likely to live in individual housing on the contrary of those from clusters 2 and 5. On Figure 4c, users from clusters 1, 4 and 6 have lower incomes, whereas travelers from clusters 2, 5 and 9 have higher incomes.

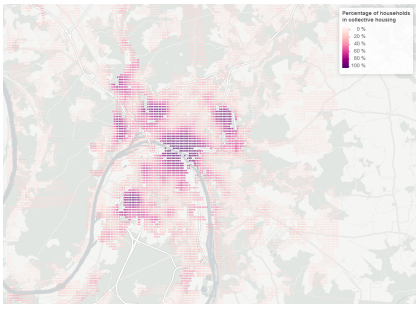
3.3 Stations profile clustering

In the case of stations clustering, it seems that results are more sensitive to the choice of H and K than for passengers clustering. It is likely linked with the low number of observations (475 stations profiles). Thus, we choose H and K in term of interpretability and setted $H = 3$ and $K = 5$.

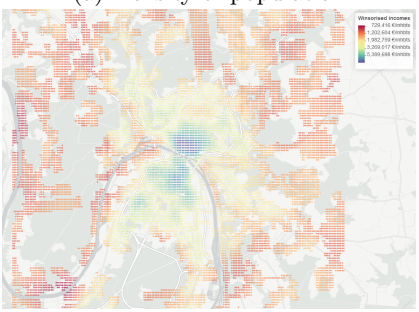
The 3 words obtained are the ones in Figure 7. The first time component is described by check-ins at 7 and 8 a.m. We will call it the “morning component”. The second time component shows check-ins at 4 and 5 p.m on Mondays, Tuesdays, Thursdays and Fridays and check-ins at 12 p.m on Wednesdays. We will



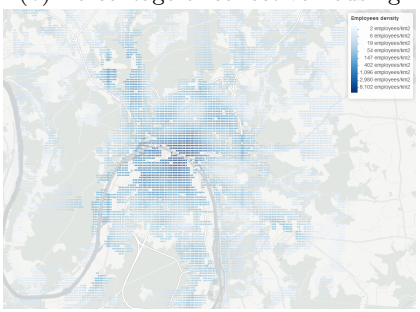
(a) Density of population



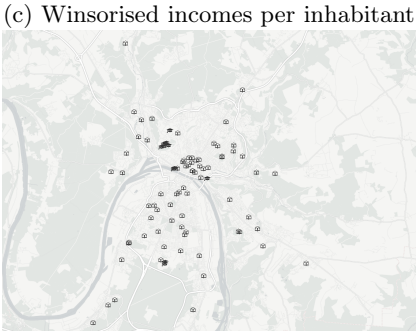
(b) Percentage of collective housing



(c) Winsorised incomes per inhabitant

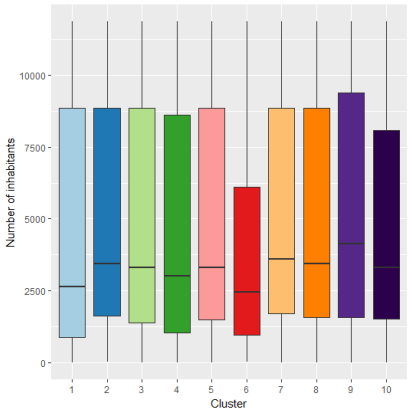


(d) Density of employees

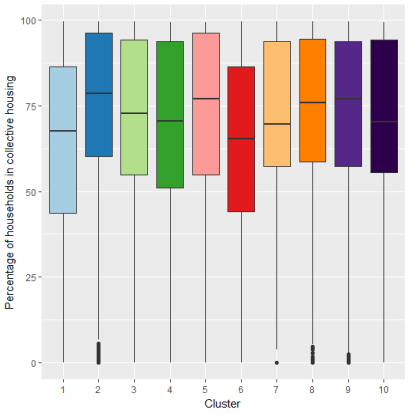


(e) Geolocation of schools and universities

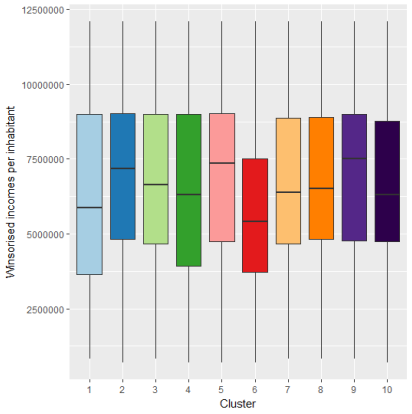
Table 3: Representation of contextual data



(a) Density of population



(b) Percentage of collective housing



(c) Winsorised incomes per inhabitant

Table 4: Contextual data through travelers clusters

Figure 7: Words obtained by NMF-EM on stations data with $K = 5$ and $H = 3$.

name it the "end of school component". The third component shows check-ins at 6 p.m, during Wednesdays afternoons, during Saturdays and off-peaks periods. This component will be called the "off-peak component".

Figure 8 shows the 5 clusters. Stations in cluster 1 are stations where there

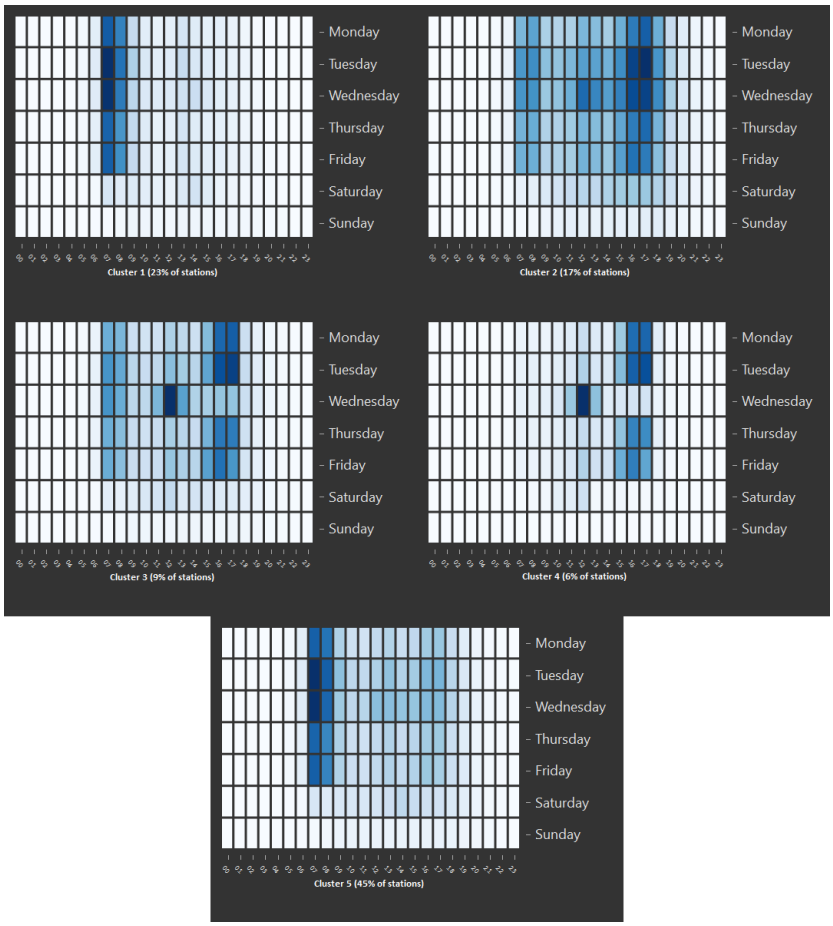
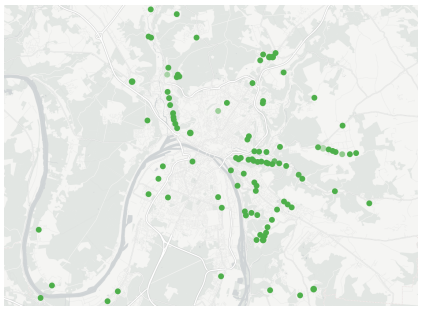


Figure 8: Clusters obtained by NMF-EM on stations data with $K = 5$ and $H = 3$.

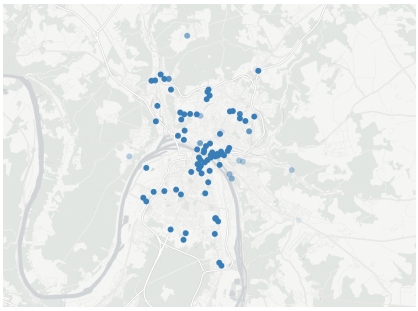
are check-ins only in the morning at 7 or 8 a.m. These stations are likely in residential areas. In cluster 2, the stations have check-ins all day long, but with highest probabilities during peaks. Stations in cluster 3 have check-ins in the morning and at the end of school. They are likely to be near schools in residentials area. Stations in cluster 4 have check-ins only at end of school times. Thus, these stations are probably near schools. Finally, stations in cluster 5 are pretty similar than the ones in cluster 1: a large majority of check-ins are made in the morning (7 or 8 p.m). The only difference is that it is more likely to have check-ins during the rest of the day in cluster 5 than in cluster 1.

The figures in Table 5 show the geographical repartition of the five clusters. In Figure 5a, we observe the stations contained in cluster 1. This cluster groups stations that have check-ins only in the morning. On the figure, we observe

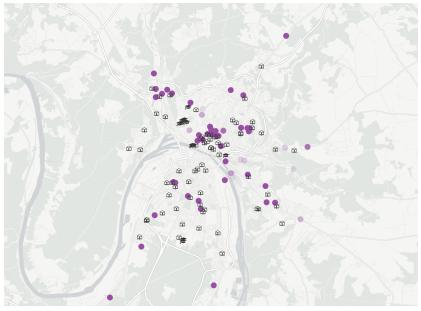
that these stations are distant from the city center and are mainly located in residential areas. Figure 5b shows stations of cluster 2, that have check-ins all day long with stronger attendance during peak-periods. These stations are mainly located in the city center. Figures 5c and 5d look alike. Indeed, clusters 3 and 4 have the "end of school" component and the points on the map are close to educational establishment. Figure 5e shows stations from cluster 5. These stations have check-ins all day long but most are made in the morning. By looking at the map, we cannot notice any significant pattern.



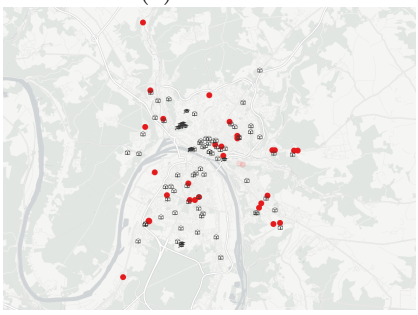
(a) Cluster 1



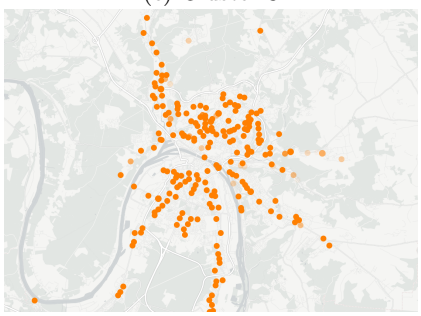
(b) Cluster 2



(c) Cluster 3



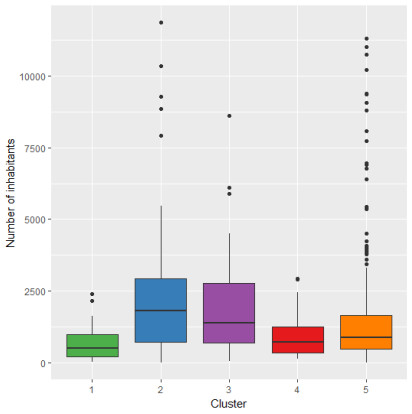
(d) Cluster 4



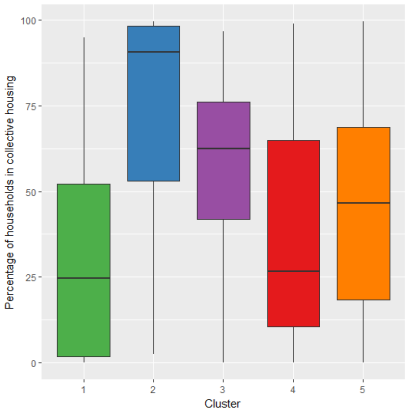
(e) Cluster 5

Table 5: Map of the stations - opacity of the points are proportional to the adequacy between the stations and the clusters.

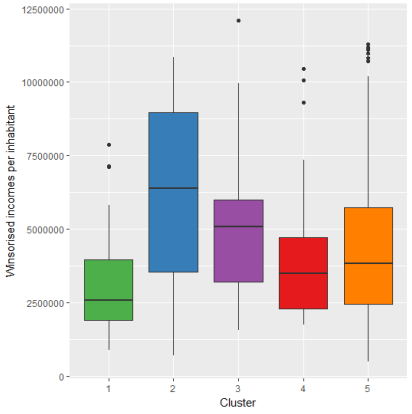
As explain above in Section 3.2, we created catchment areas around the stations. Thanks to these areas, we can obtain a description of the neighborhood around each of them. Results are presented in Table 6.



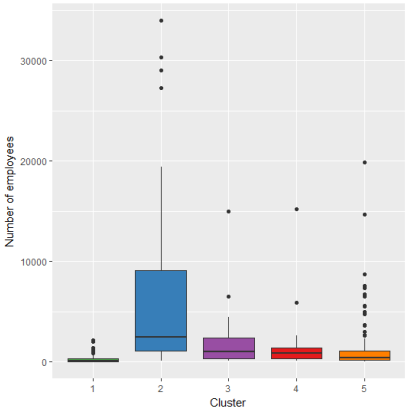
(a) Number of inhabitants



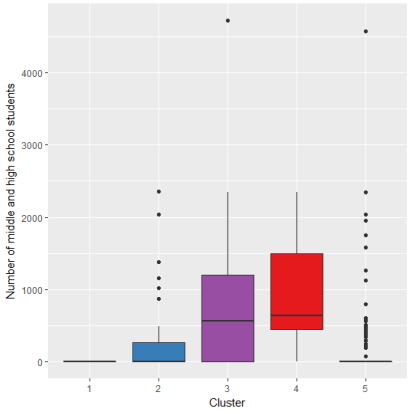
(b) Percentage of households in collective housing



(c) Winsorised incomes per inhabitant



(d) Number of employees



(e) Number of schoolkids

Table 6: Description of the stations clusters through contextual data

We observe in Figure 6a that stations of clusters 2 and 3 are more likely to attract

more inhabitants around them. In Figure 6b, we note that stations in clusters 2 and 3 are located in areas with more collective housing, whereas clusters 1 and 4 are in areas with more individual housing. We notice in Figure 6c that stations in cluster 2 have more variability in the incomes of their neighborhoods. Cluster 1 is the cluster with the lowest incomes around its stations. In Figures 6d, we observe that stations from cluster 2 are more located in activity areas than the others. Finally, we notice in Figure 6e that stations from clusters 3 and 4, as seen on figures 5c and 5d, are catching lot of schoolkids (11 to 18 years old) since they are located near middle and highschools.

To conclude, we can say that stations from cluster 1 are located in peripheral, where there are few inhabitants, housing are mostly individual and incomes are low. Stations from cluster 2 are mostly located in the city center where the population is high, have high incomes and lives in collective housing and where companies are numerous. Stations from clusters 3 and 4 are located near middle and high schools. However, stations from cluster 3 are most located in rich collective housing areas than stations from cluster 4. As cluster 5 contains 45% of the stations, the indicators are on average.

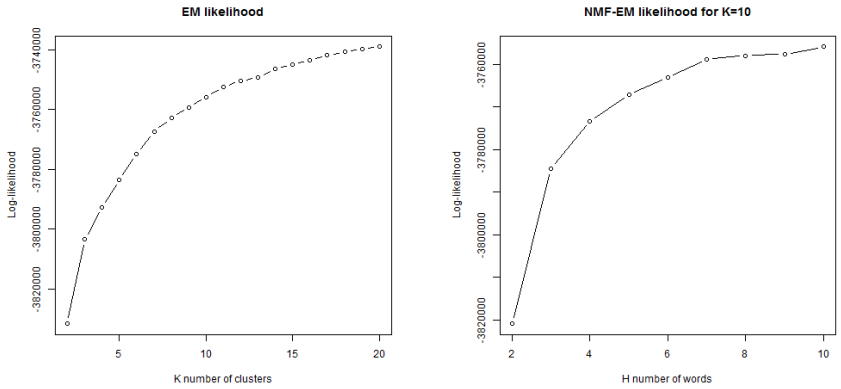
3.4 Passengers profile clustering on another network

Since customs and habits are different from one country to another, we applied the NMF-EM algorithm on another Transdev network located in the Netherlands. Validation data used are the ones from November 2015, because it is one of the months with the most typical behaviours on the network. During this month, $1,743,574$ check-ins have been made by $200,429$ travelers. For the same reasons as those mentionned above, we kept only the users who have been travelling on the network minimum 4 days during the study period and who have made their first boarding after 4 a.m each day at the same station at least 50% of the time. We ended up with $55,149$ regular card holders, that made $1,245,011$ check-ins.

By using the same model selection method as in Section 3.2 (i.e. slope heuristic), we obtained the optimal values of $K = 10$ and $H = 7$ (cf. Figures 7a and 7b in Table 7).Corresponding words and clusters are contained respectively in Figures 9 and 10.

The interpretation of the words is:

1. Word 1: travels at 6 or 7 a.m and slightly around 4 p.m during the week.
2. Word 2: travels during the week-end.
3. Word 3: diffuse travel habits from 8 a.m to 4 p.m Mondays to Fridays.
4. Word 4: travels at 7a.m on weekdays.
5. Word 5: diffuse habits with highest probabilities from 5 p.m to 12 a.m during the week.
6. Word 6: diffuse habits from 9 a.m to 5 p.m with highest probability at 1 p.m Mondays to Saturdays.



(a) Log-likelihood as a function of K, under $H = K$

(b) Log-likelihood as a function of $H \in \{2, ..., K\}$ under $K = 10$

Table 7: Model selection for the dutch network

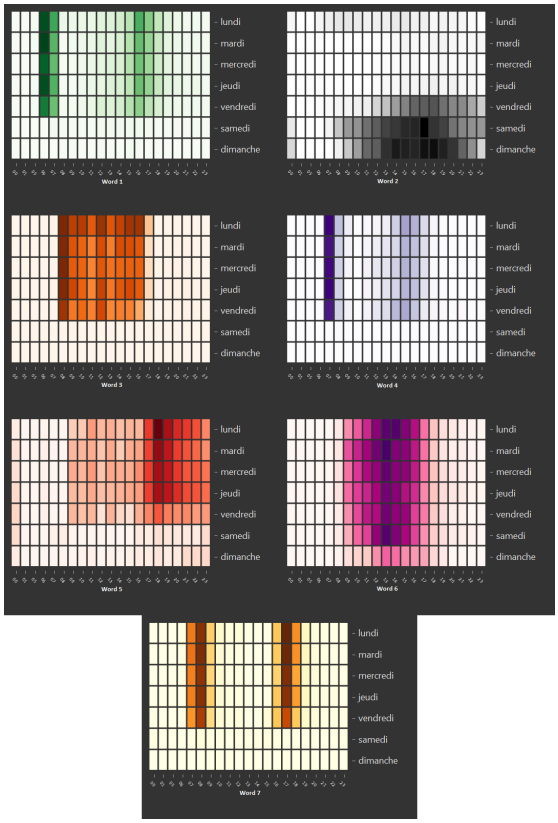


Figure 9: Words obtained by NMF-EM on users data with $K = 10$ and $H = 7$.

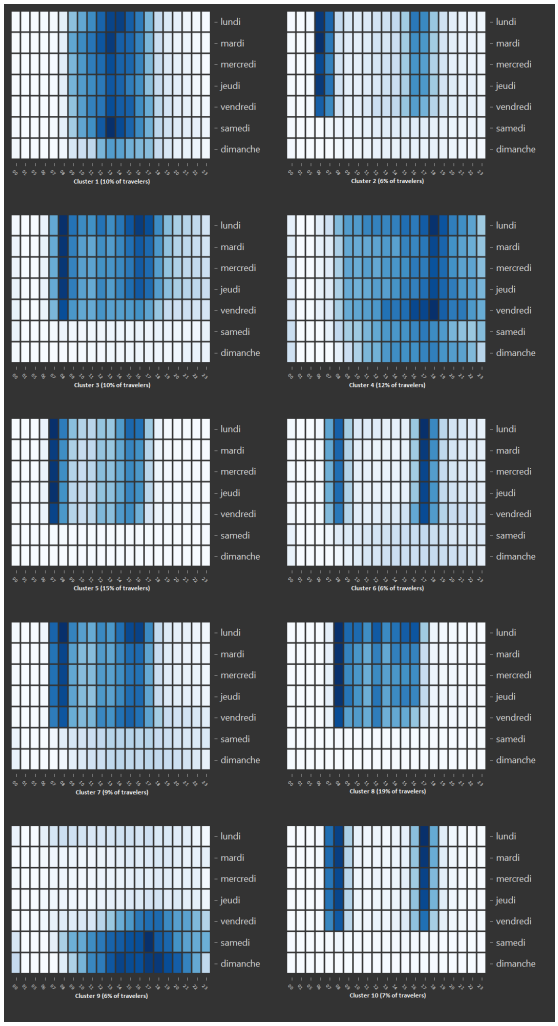


Figure 10: Clusters obtained by NMF-EM on users data with $K = 10$ and $H = 7$.

7. Word 7: travels at 8 a.m and 5 p.m.

We can interpret the cluster as follows:

1. Cluster 1: diffuse habits from 9 a.m to 5 p.m with highest probability at 1 p.m Mondays to Saturdays.
2. Cluster 2: travels at 6 or 7 a.m and at 4 or 5 p.m during the week.
3. Cluster 3: diffuse habits from 7 a.m to 6 p.m on weekdays.
4. Cluster 4: diffuse travel habits from 9 a.m to 11 p.m.
5. Cluster 5: travels at 7 or 8 a.m diffuse habits during the afternoon.
6. Cluster 6: travels at 8 a.m and 5 p.m.
7. Cluster 7: diffuse travel habits from 7 a.m to 5 p.m Mondays to Fridays.
8. Cluster 8: diffuse habits from 8 a.m to 4 p.m during the week.
9. Cluster 9: travels during the week-end.
10. Cluster 10: travels at 7 or 8 a.m and around 4 p.m.

As for the french network, the NMF-EM algorithm allowed to identify group of travelers that have similar temporal habits within them but very distinct between them.

4 Conclusion

In this paper, using smart card data allowed us to obtain a new method to cluster passengers. In addition to highlighting typical temporal profiles, we insisted on temporal components that enable more precise knowledge on travelers temporal habits.

Secondly, we used data contained in the product used to get a description of the travelers. Indeed, the age ranges and ticket type added informations on proifiles, such that their type of activities (unemployed, students, retired, working population,...).

Finally, adding socioeconomic data finished to draw a portrait of typical users. Especially by using data on home neighbourhood, like the number of employees or type of housing.

Without any personnal data, couple smart card and socioeconomic data permits to obtain rich sources of information on typical profiles from network users. This could allow networks to have a better customer knowledge, and adapt their price offer to the profiles described.

In Netherlands, not only check-ins, but although check-outs are recorded. The NMF-EM algorithm could be therefore applied to the number of minutes spent in a public transport system by time slot.

Acknowledgements

The authors would like to thank Denis COUTROT and Nadir MEZIANI from Transdev for their support and comments on previous versions of this work.

References

[1] P. Alquier and B. Guedj. An oracle inequality for quasi-bayesian non-negative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017.

[2] C. Biernacki, G. Celeux, and G. Govaert. An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272, 1999.

[3] C. Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[7] L. Carel and P. Alquier. Non-negative matrix factorization as a pre-processing tool for travelers temporal profiles clustering. In M. Verleysen, editor, *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 417–422. i6doc.com, 2017.

[8] E. Côme and L. Oukhellou. Model-based count series clustering for bike sharing system usage mining: a case study with the vélib'system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):39, 2014.

[9] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[10] M. K. El Mahrsi, E. Côme, J. Baro, and L. Oukhellou. Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. In *ACM SIGKDD Workshop on Urban Computing*, 2014.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[12] M. Fop and T. B. Murphy. Variable selection methods for model based clustering. *arXiv preprint arXiv:1707.00306*, 2017.

[13] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[14] E. F. Gonzalez and Y. Zhang. Accelerating the lee-seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, 2005.

[15] R. Hamon, P. Borgnat, C. Févotte, P. Flandrin, and C. Robardet. Factorisation de réseaux temporels: étude des rythmes hebdomadaires du système vélo'v. In *Colloque GRETSI 2015*, 2015.

[16] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[17] D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[18] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[19] C.-J. Lin. Projected Gradient Methods for Non-negative Matrix Factorization. *Neural computation*, 19(10):2756–2779, 2007.

[20] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[21] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009.

[22] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[23] J. Mei, Y. De Castro, Y. Goude, and G. Hébrail. Recovering multiple nonnegative time series from a few temporal aggregates. *arXiv preprint arXiv:1610.01492*, 2016.

[24] C. Morency, M. Trepanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[25] J. W. Paisley, D. M. Blei, and M. I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference., 2014.

[26] M.-P. Pelletier, M. Trépanier, and C. Morency. *Smart card data in public transit planning: a review*. CIRRELT, 2009.

[27] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò. Collective human mobility pattern from taxi trips in urban area. *PloS one*, 7(4):e34487, 2012.

[28] M. Poussevin, E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Mining ticketing logs for usage characterization with nonnegative matrix factorization. In *International Workshop on Modeling Social Media*, pages 147–164. Springer, 2014.

[29] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[30] A. N. Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert. Clustering the velib origin-destinations flows by means of poisson mixture models. In *ESANN*, 2013.

[31] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[32] D. Steinley and M. J. Brusco. Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73(1):125–144, 2008.

[33] D. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6201–6205. IEEE, 2014.

[34] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[35] Y. Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[36] Z. Yang, J. Corander, and E. Oja. Low-rank doubly stochastic matrix decomposition for cluster analysis. *Journal of Machine Learning Research*, 17(187):1–25, 2016.

[37] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

A Users location

(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Table 8: Share of clusters per home station - Clusters 1 to 6

(a) Cluster 7

(b) Cluster 8

(c) Cluster 9

(d) Cluster 10

Table 9: Share of clusters per home station - Clusters 7 to 10