Rapport d'étude

Etude du jeu de données MovieLens

Analyse descriptive du comportement des groupes d'utilisateurs

L3
CMI SID

Max Halford
Giovanni Zanitti

Professeur encadrant
Jonathan Louëdec

Contents

1 Objectif **1**

2 Choix des données et munging **2**

3 Analyses Factorielles des Composantes **5**
3.1 Genre(s) des films — Catégorie d'âge des utilisateurs 5
3.1.1 Femmes 5
3.1.2 Hommes 6
3.2 Genre(s) des films — Occupation professionnelle des utilisateurs 8
3.2.1 Femmes 8
3.2.2 Hommes 8
3.3 Genre(s) des films — Région d'habitation des utilisateurs . . . 8
3.3.1 Femmes 8
3.3.2 Hommes 8
3.4 Décennie de sortie des films — Catégorie d'âge des utilisateurs 8
3.4.1 Femmes 8
3.4.2 Hommes 8
3.5 Décennie de sortie des films — Occupation professionnelle des utilisateurs 8
3.5.1 Femmes 8
3.5.2 Hommes 8
3.6 Décennie de sortie des films — Région d'habitation des utilisateurs 8
3.6.1 Femmes 8
3.6.2 Hommes 8

4 Classification **13**

5 Conclusion **14**

Chapter 1

Objectif

Le but de l'analyse est de répondre à la question "**Qui aime quoi?**". Le résultat sera une liste d'affirmations, par exemple, "*Les hommes qui ont la trentaine aiment les films d'actions*".

Il est important de remarquer qu'une affirmation peut-être:

- Plus ou moins vraie
- Plus ou moins détaillée

Evidemment seulement les affirmations qui sont nettement visibles seront gardées. Il n'est pas dit que des groupes d'utilisateurs existent. Au contraire le fait qu'il n'y en ait pas est aussi une information.

L'intuition devrait nous dire qu'on peut trouver des groupes plus ou moins précis. Le but est de trouver des groupes généralistes, peut-être dans l'optique de faire de la recommandation. Par exemple est-il plus intéressant de savoir que 70 pourcent des hommes du Midwest aiment les films d'actions que de savoir que 90 pourcent des avocats hommes du Midwest ayant la trentaine aiment les films d'action-aventure? Sachant que le deuxième groupe est plus restreint que le premier.

L'étude se fait en trois parties, d'abord des Analyses Factorielles des Correspondances (AFC) sont produites pour observer les liens entre deux variables et ainsi déduire des liens globaux. Ensuite une Analyse des Correspondances Multiples est réalisée pour extraire des liens plus détaillées. Enfin on opère une classification des utilisateurs pour lister des groupes selon les préférences.

Chapter 2

Choix des données et munging

Tout d'abord **seulement les notes de 4 et 5 sont conservées**. En effet le but est de connaitre les préférences des utilisateurs, pas les genres de films qu'ils notent.

Ensuite il s'agit de choisir les variables à étudier. Pour les utilisateurs sont gardés:

- L'âge
- Le sexe
- L'occupation professionnelle
- Le code ZIP

Le munging est comme suit: l'âge devient une variable qualitative en ne conservant que le premier chiffre de l'âge (35 devient 30, 42 devient 40 etc.). Puisque le sexe est variable binaire, les utilisateurs sont scindés en deux groupes pendant les AFCs, en effet cela permet de rajouter une variable aux AFC qui étudient les relations entre deux variables. L'occupation professionnelle reste intacte. Chaque code ZIP est associé à l'état américain qui lui est propre ou bien à la valeur *Other* pour les codes éronnés et les pays étrangers aux Etats-Unis.

Pour les films sont gardés:

- La date de sortie
- Le(s) genre(s)

Pour le munging la décennie de sortie du film est extraite. Un film peut avoir plusieurs genres et donc un choix s'impose. Pour les AFCs les tableaux de contingence prennent en compte le fait qu'aimer un film qui a deux genres équivaut à aimer un film du premier genre et un autre film d'un autre genre. Pour l'ACM on procède de la même façon.

Tout le munging est effectué en *Python* avec l'aide du module *pandas*. Une jointure est effectuée sur les trois fichiers csv fournis. Des tableaux de contingence pour les AFCs sont sauvegardés sous format csv, les analyses sont effectués avec *R*. Les données sont filtrées comme indiqué précédemment. Les mêmes tableaux de contingence sont construits pour les hommes et les femmes. L'API *api.zippopotam.us* est utilisée pour convetir les codes ZIP en états américains qui sont ensuite filtrée selon la carte suivante. Toutes les

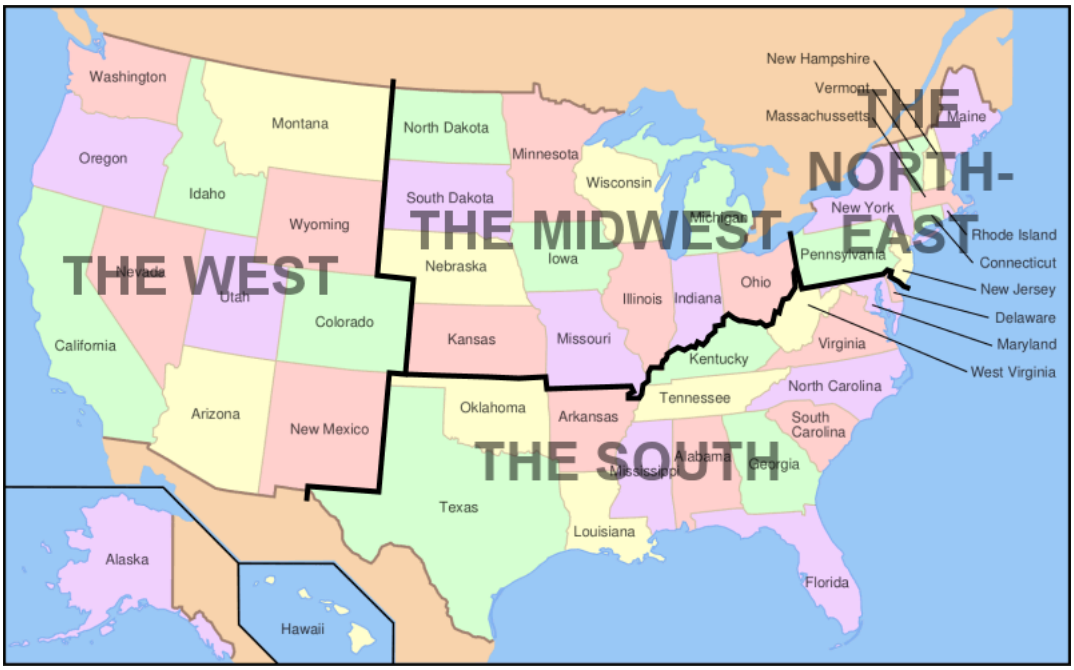

Figure 2.1: Régions des Etats-Unis

données sont sauvegardées comme suit:

- Un fichier csv contenant toutes les données.
- Une fichier csv contenant les données utiles à l'ACM.
- Une base de donnée MongoDB contenant toutes les données.

Le premier fichier sert à construire les tableaux de contingence pour produire les AFCs. Le deuxième diffère du premier car il ne contient pas toutes les données, de plus les genres de films sont séparés (par exemple un film ayant trois genres devient trois lignes). La base MongoDB est utilisée pour deux raisons. D'une part la création des tableaux de contingence pour les AFCs relatifs aux genres des films est très rapide, en effet le module *pymongo* renvoit un curseur Python, un type de données natif qui permet d'itérer plus rapidement que sur un tableau de données. D'autre part une base de données peut servir dans le cas de l'étude d'une sous-population particulière d'une grande population, en effet les temps de calculs pourraient être long si on élargit notre étude au jeu de données de 10 millions de lignes de MovieLens. De la même façon calculer l'ACM avec *R* pourra se révéler hardu sur le jeu de données de 10 millions de ligne, pour cela il serait envisageable d'utiliser *Julia*, notamment grâce à https://github.com/MaxHalford/MultivariateStats.jl.

Chapter 3

Analyses Factorielles des Composantes

3.1 Genre(s) des films — Catégorie d'âge des utilisateurs

3.1.1 Femmes

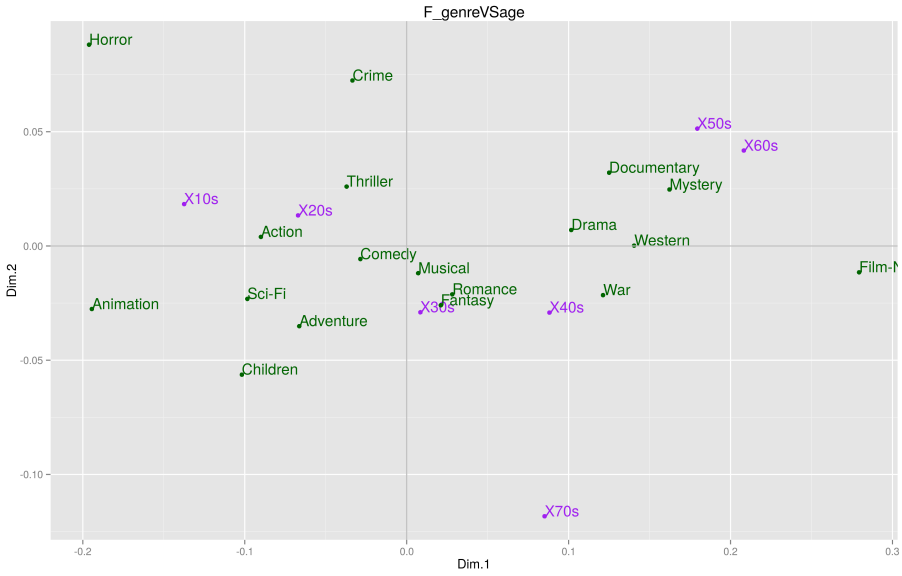

Figure 3.1: Genres des films par rapport à l'âge des utilisatrices

3.1.2 Hommes

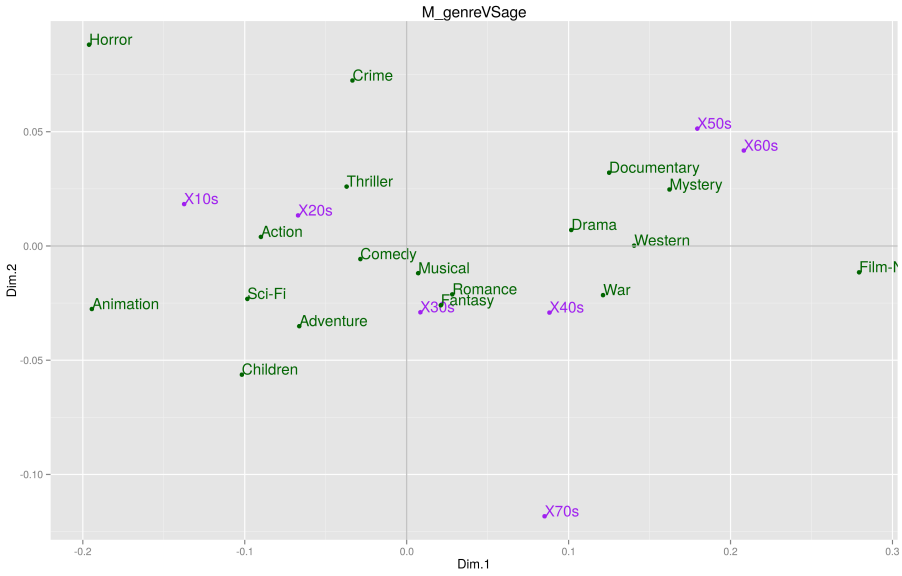

Figure 3.2: Genres des films par rapport à l'âge des utilisateurs

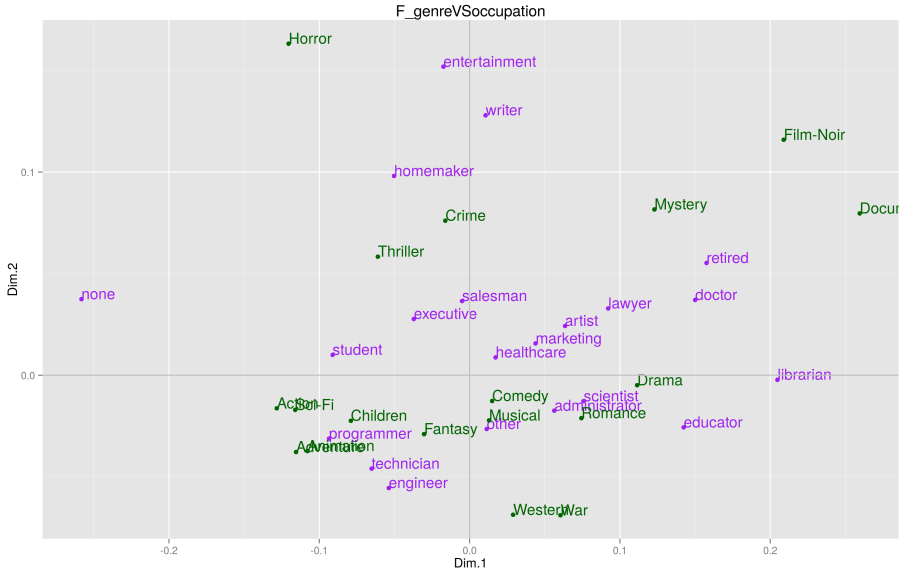

Figure 3.3: Genres des films par rapport à l'occupation professionnelle des utilisatrices

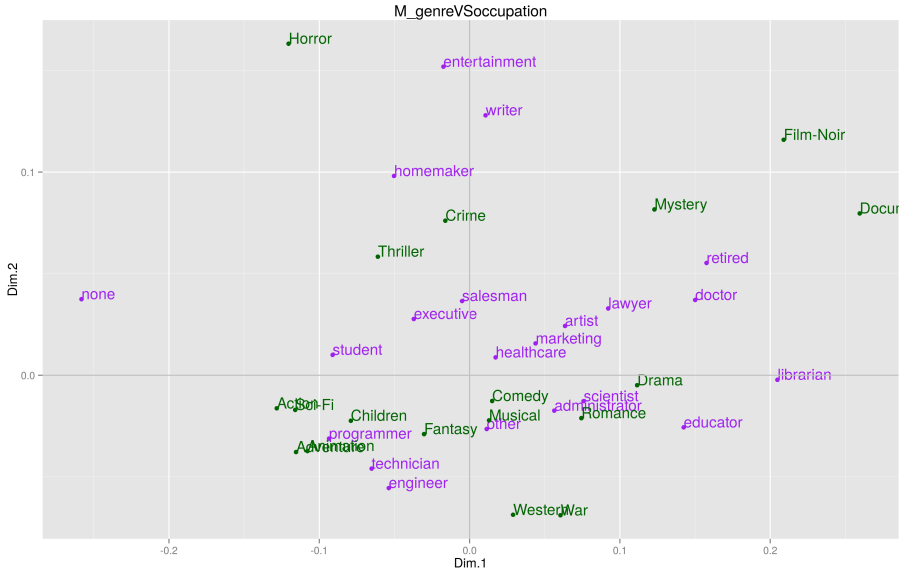

Figure 3.4: Genres des films par rapport à l'occupation professionnelle des utilisateurs

3.2 Genre(s) des films — Occupation professionnelle des utilisateurs

3.2.1 Femmes

3.2.2 Hommes

3.3 Genre(s) des films — Région d'habitation des utilisateurs

3.3.1 Femmes

3.3.2 Hommes

3.4 Décennie de sortie des films — Catégorie d'âge des utilisateurs

3.4.1 Femmes

3.4.2 Hommes

3.5 Décennie de sortie des films — Occupation professionnelle des utilisateurs

3.5.1 Femmes

3.5.2 Hommes

3.6 Décennie de sortie des films — Région

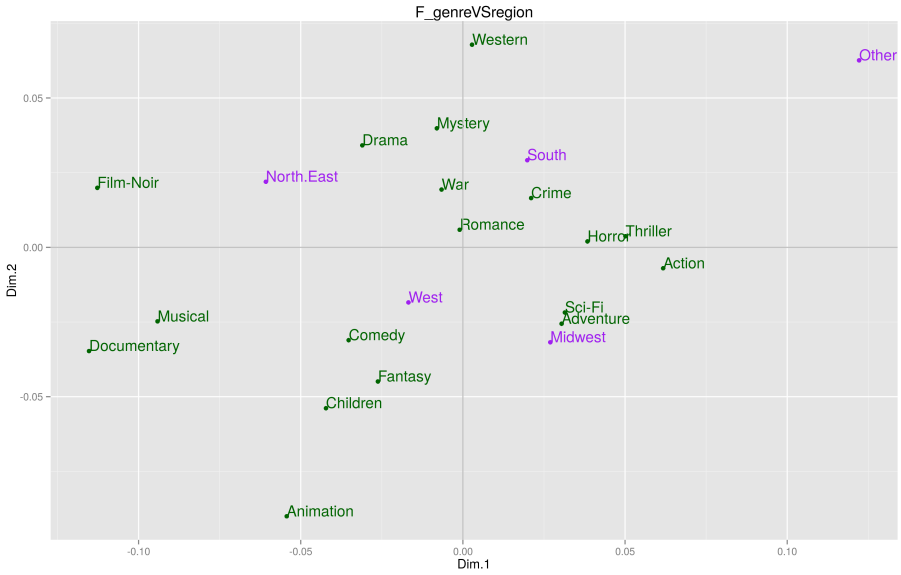

Figure 3.5: Genres des films par rapport à la région d'habitation des utilisatrices

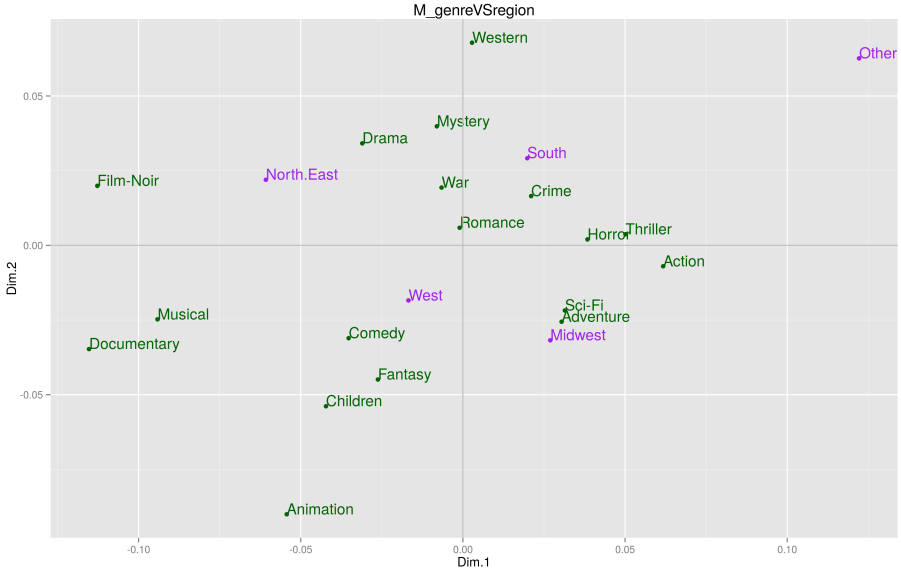

Figure 3.6: Genres des films par rapport à la région d'habitation des utilisateurs

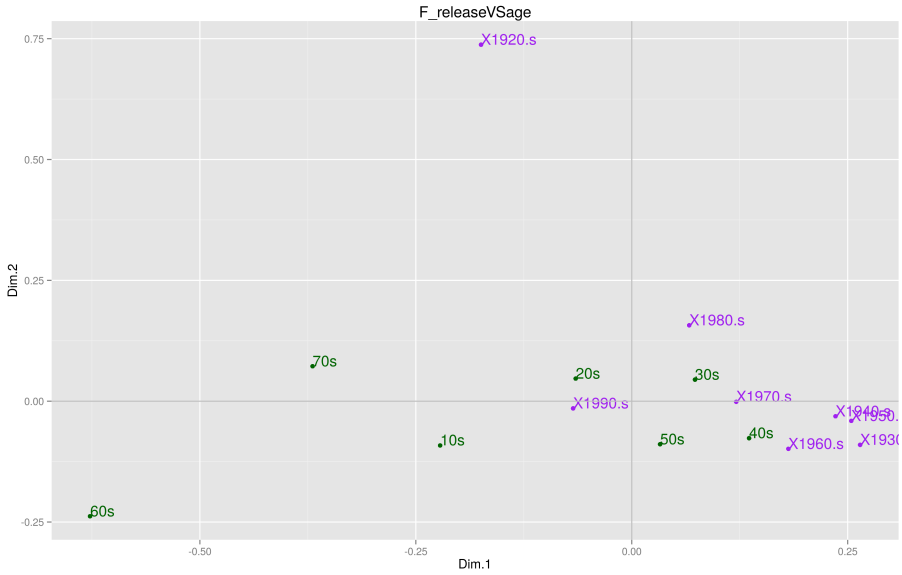

Figure 3.7: Décennie de sortie des films par rapport à l'âge des utilisatrices

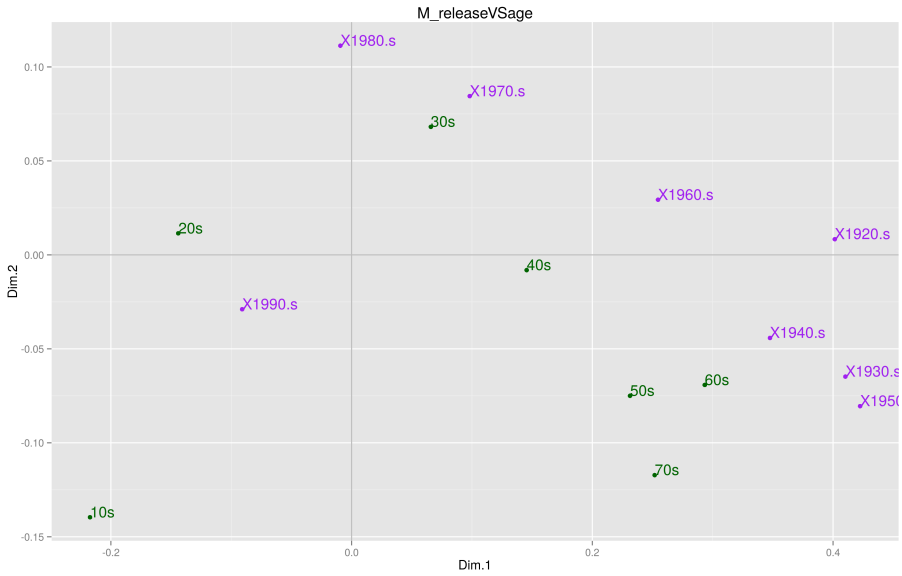

Figure 3.8: Décennie de sortie des films par rapport à l'âge des utilisateurs

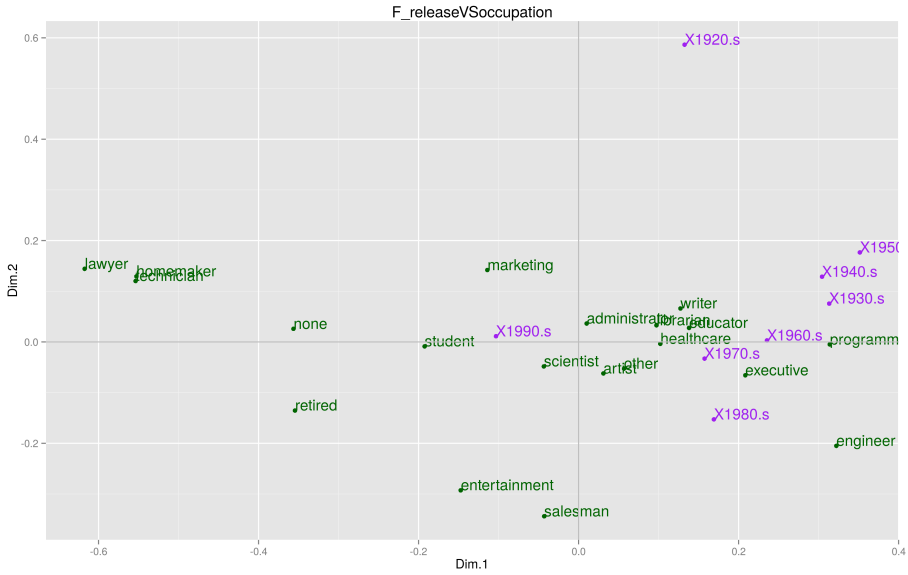

Figure 3.9: Décennie de sortie des films par rapport à l'occupation professionnelle des utilisatrices

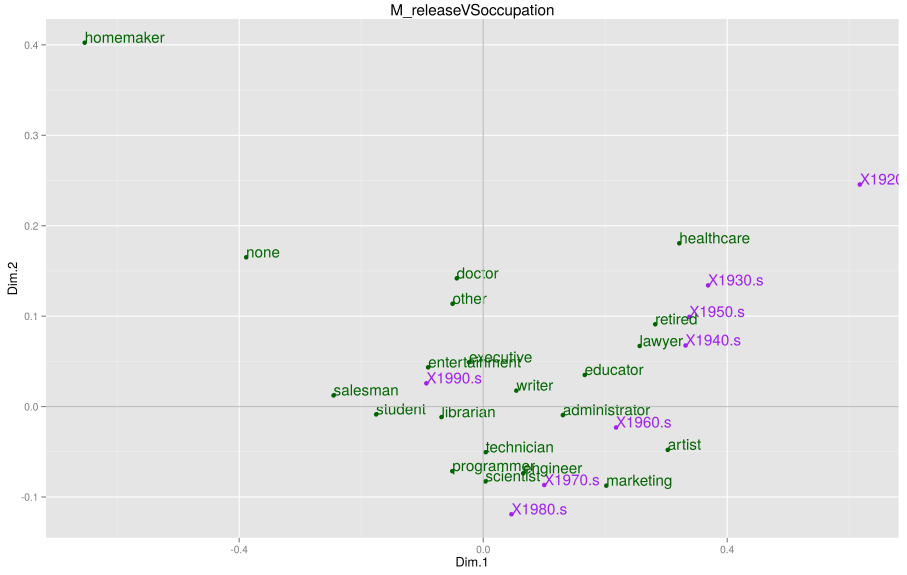

Figure 3.10: Décennie de sortie des films par rapport à l'occupation professionnelle des utilisateurs

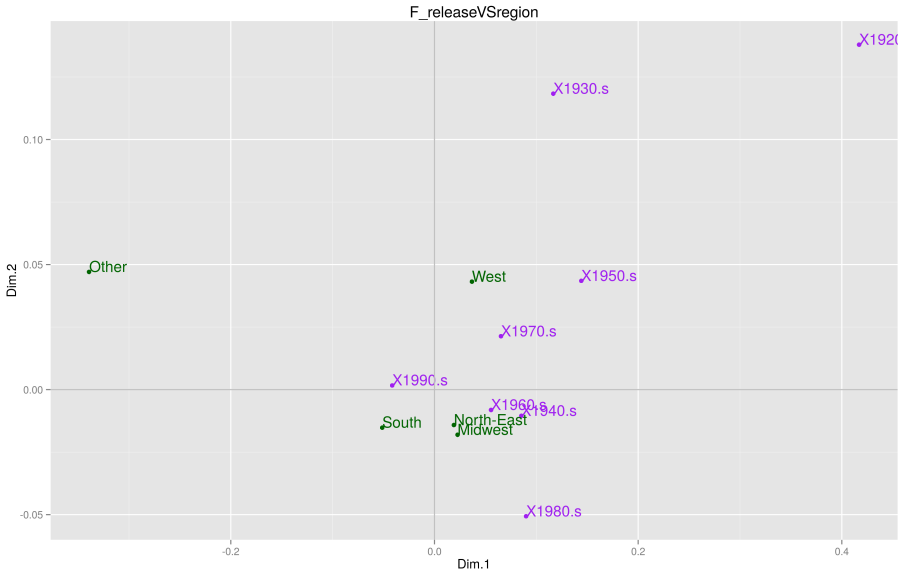

Figure 3.11: Décennie de sortie des films par rapport à la région d'habitation des utilisatrices

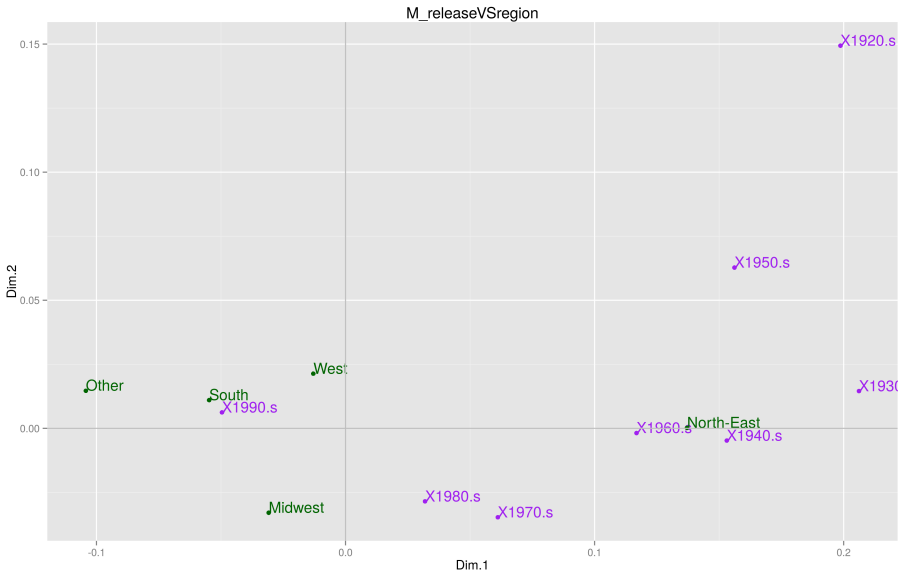

Figure 3.12: Décennie de sortie des films par rapport à la région d'habitation des utilisateurs

Chapter 4

Classification

Chapter 5

Conclusion

¡Conclusion here¿