

# Rapport technique SIMBA

## *Application pour la fouille et l'analyse de séquences sémantiques*

\* \*

### Table des matières

1. Présentation du projet Simba .....	2
2. Chargement de fichiers avec Simba .....	2
2.1 Chargement des fichiers principaux.....	2
2.2 Ontologies des activités.....	5
3. Filtrage de données.....	6
4. Statistiques globales .....	8
4.1 Statistiques sur les séquences.....	9
4.1.1 Distribution du nombre d'activités journalières.....	9
4.1.2 Distribution globale des activités .....	9
4.1.3 Diagramme de flots (ou Chord Diagram) .....	10
4.1.4 Motifs topologique (ou Daily pattern) .....	12
4.1.5 Tapis de séquences .....	14
4.2 Statistiques sur les individus.....	15
5. Clustering.....	16
5.1 Qualité des clusters.....	17
5.2 Barres empilées et analyse d'activités par cluster .....	18
5.3 Diagramme mosaïque par cluster .....	19
5.4 Nuages et ciel de mots .....	19
6. Conclusions et perspectives .....	21

## 1. Présentation du projet Simba

SIMBA (pour *Sematic Mobility Behavior Analysis*) est une application web développée en R Shiny basée sur une méthodologie de découverte de connaissances intégrant de nombreux indicateurs et techniques pour l'analyse de séquences sémantiques.

La conception de SIMBA a été pensée en collaboration avec des experts métiers issus du collectif MOBI'KIDS et à fait l'objet d'un stage de développement de Master 2. Bien qu'imaginée initialement pour permettre l'analyse des données issues du projet MOBI'KIDS, SIMBA se veut être une plate-forme générique pour l'exploration interactive de tout type de séquence de mobilité sémantique.

Le rapport est structuré autour des fonctionnalités de SIMBA. La **section 2** aborde le chargement des fichiers (données des séquences, individus et matrice de distance) ainsi que des ontologies qui viennent enrichir et structurer les activités au sein des séquences sémantiques.

La **section 3** présente les fonctionnalités de filtrage et de mise à jour des données. La **section 4** est dédiée à l'analyse statistiques des données. Une première sous-section détaille les différents indicateurs disponibles pour l'analyse des séquences ; une seconde sous-section aborde l'analyse bivariée sur les données des individus.

La **section 5** présente les fonctionnalités liés au clustering de séquences et les différentes analyses par cluster pouvant être réalisés.

Le rapport se conclut par la **section 6** qui résume l'état actuel de l'application ainsi que les perspectives et améliorations futures pouvant être apportées au prototype SIMBA.

## 2. Chargement de fichiers avec Simba

### 2.1 Chargement des fichiers principaux

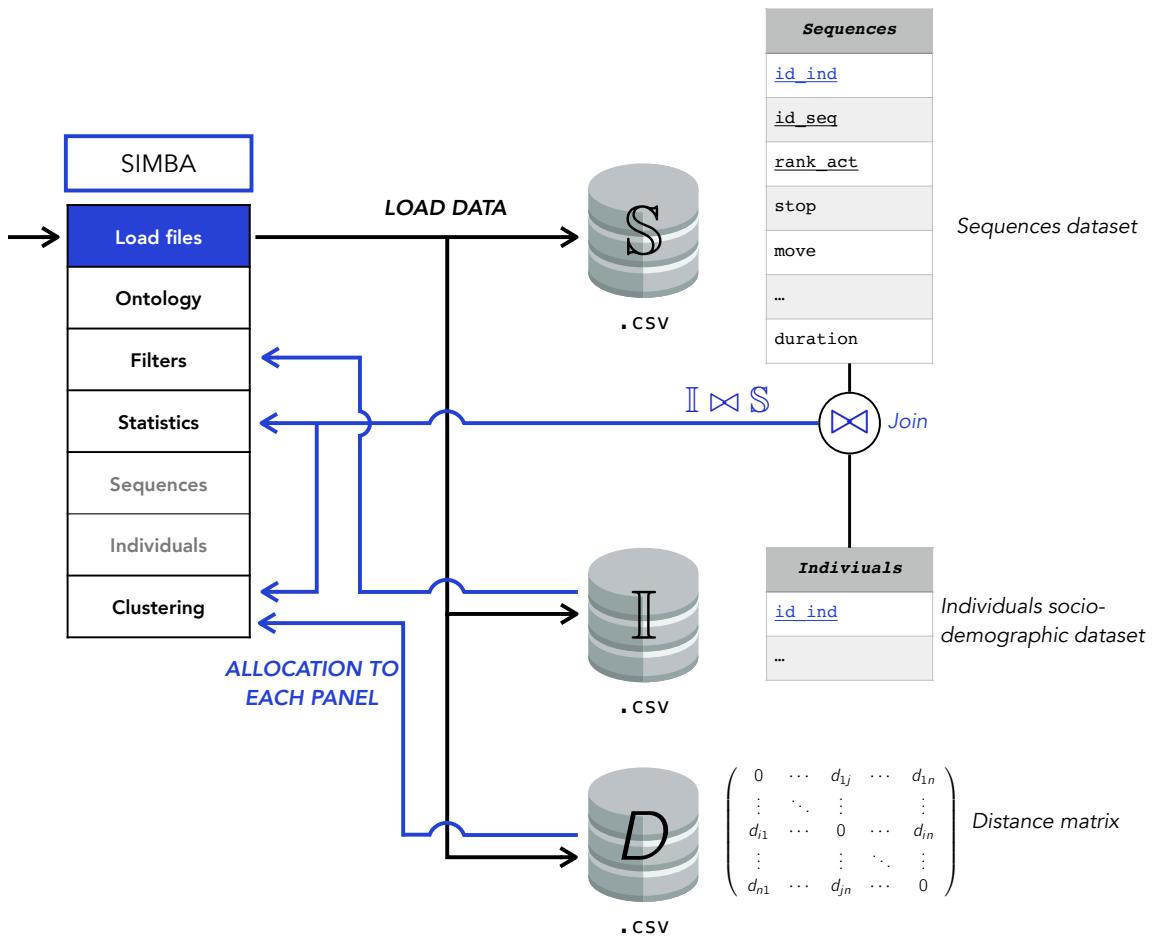
Dans une recherche de simplicité et de minimalisme, l'application SIMBA a été conçue selon un ensemble de **cinq panels principaux** permettant la navigation et l'interaction avec les données :

- Chargement des données (*Load files*)
- Chargement et visualisation des ontologies (*Ontology*)
- Filtrage des données (*Filters*)
- Statistiques générales (*Statistics*), dotée de deux sous onglets :
  - Sur les séquences (*Sequences*)
  - Sur les individus (*Individuals*)
- Fouille et clustering interactif (*Clustering*)

Une vue globale de l'architecture est donnée par la figure ci-dessous. Les cinq panels principaux sont alimentés par **trois fichiers de données au format .csv** nécessaires au bon fonctionnement de SIMBA :

- L'ensemble des séquences (noté  $\mathbb{S}$ )

`Sequences(id_ind, id_seq, rank_act, stop, move, ..., duration)`



#### Chargement des fichiers dans SIMBA et allocation

Le couple (`id_seq`, `rank_act`) fournit la clé primaire de la table et désigne respectivement l'identifiant de séquence et rang de l'activité dans la séquence. La colonne `id_ind` renseigne l'identifiant de l'individu ayant réalisé la séquence.

Le symbole `...` indique que le schéma peut être complété par des dimensions supplémentaires optionnelles (ex. lieux, accompagnement, etc.). La colonne `duration` indique la durée totale de l'activité. La table suivante présente un exemple de données pour trois

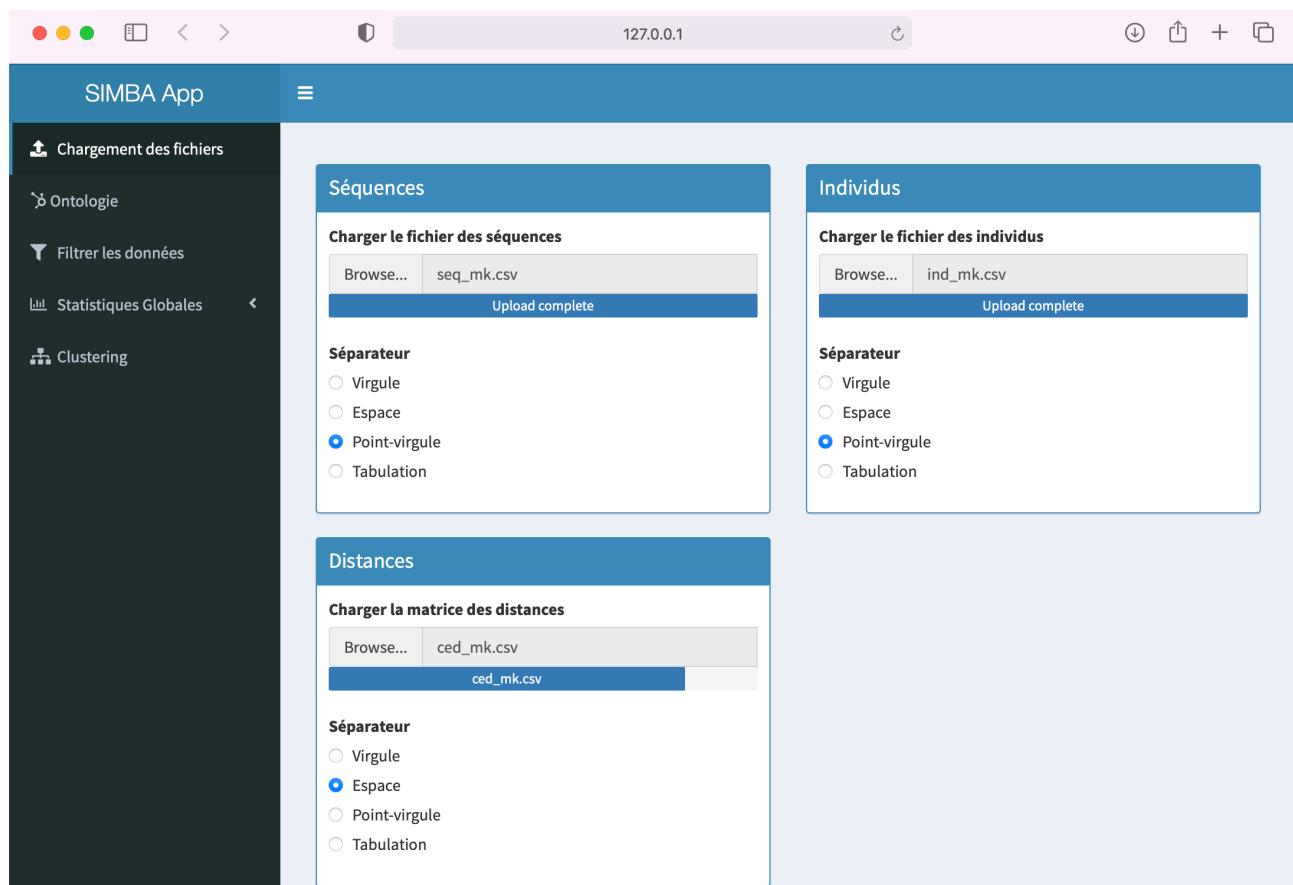
<code>id_ind</code>	<code>id_seq</code>	<code>rank_act</code>	<code>stop</code>	<code>move</code>	<code>duration</code>
1	1	1	11 12 631		600
1	1	2		B221	210
1	1	3	11 12 13		630
2	2	1	11 12 13		240
2	2	2		B31 B34	15
2	2	3	2		20
2	2	4		B221	20
2	2	5	11 12 13		1145
1	3	1	11 12 13		1400

#### Exemple de données de la table Sequences \$S\$

séquences fictives décrites sur 24h (la colonne **duration**) est en minutes. Les numéros dans les colonnes **stop** et **move** correspondent à des identifiants issus des ontologies.

- L'ensemble des informations socio-démographiques sur les individus (noté  $\mathbb{I}$ ). La table *Individuals* possède la clé primaire `id_ind` permettant d'effectuer une jointure avec la table *Sequences*. Hormis cette clé, la table ne possède pas de colonne imposée.
- La matrice de distance  $D$  entre chacune couple de séquences est décrite selon une représentation matricielle à 2 dimensions. Celle-ci est pré-calculée et chargée par l'utilisateur pour éviter des temps de calcul potentiellement importants.

Le chargement est la première étape de l'application SIMBA. Il s'agit aussi de la page d'accueil de l'application. La figure précédente présente le rendu visuel de l'application.



Page de chargement des fichiers de SIMBA

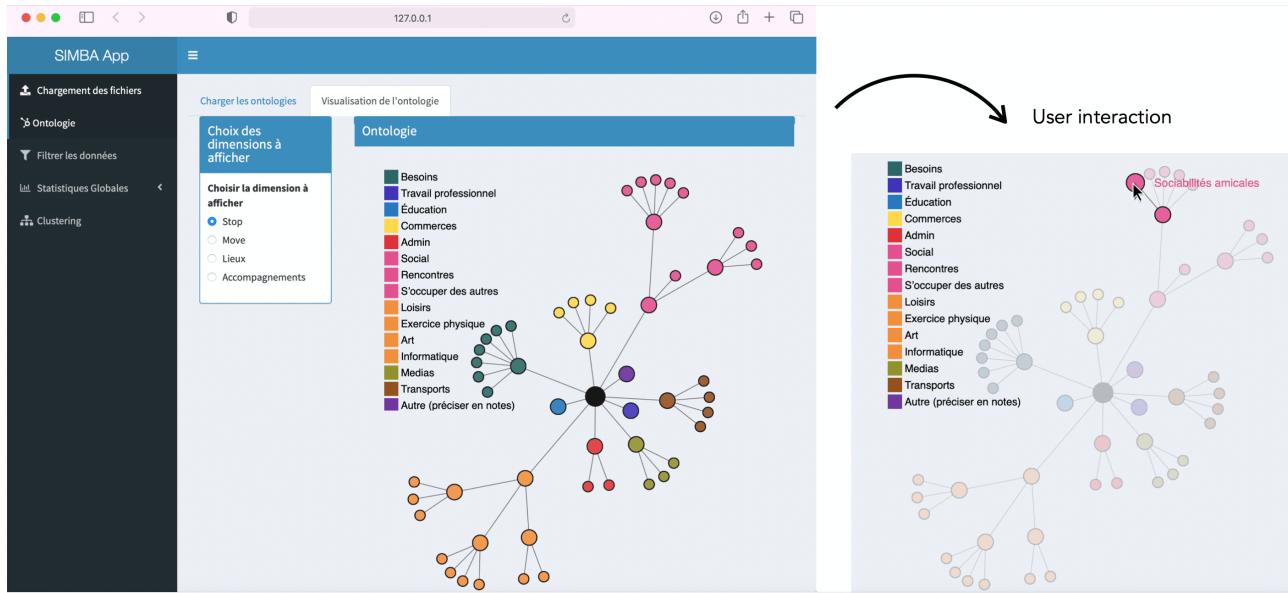
### Point technique

Le chargement des fichiers et l'implémentation de la page de chargement sont assurés par les fichiers :

- **simba\_ui\_load.R**  
Design et rendu visuel de la page
- **simba\_server.R**  
Stockage et jointure des fichiers chargés.

## 2.2 Ontologies des activités

SIMBA permet d'organiser les activités référencées au sein du fichier des séquences sémantiques (**Sequences.csv**) au sein d'une taxonomie / graphe de connaissances.



Visualisation d'ontologie dans SIMBA

Cette hiérarchisation (utilisée à l'origine pour le calcul des distances entre séquences) permet à l'application de regrouper certains concepts en une unique catégorie (e.g., Musique dans la métacatégorie Art) afin d'adapter et d'alléger la visualisation d'information.

Les ontologies sont représentées par la liste des arcs (`id ← parent`) sous le format d'entrée (fichier **.csv**) suivant :

id	name	parent	couleur
7	Loisirs	All	#F28F3D
71	Exercice physique	7	#F28F3D
711	Activité physique	71	#F28F3D
712	Jeu extérieur	71	#F28F3D
713	Activité plein air	71	#F28F3D
72	Art	7	#F28F3D
721	Musique	72	#F28F3D
722	Activité culturelle	72	#F28F3D

Exemple de données d'ontologie

La colonne **couleur** renseigne la couleur apparaissant sur les graphiques pour l'activité désignée.

Une ontologie doit être assignée par dimension (stop, move, ...) de la table **Sequences**.

## Point technique

Le chargement des fichiers, la visualisation des ontologies et l'organisation des concepts sont assurés par les fichiers :

- **simba\_ui\_ontology.R**  
Design de la page
- **simba\_server\_ontology.R**
  - Chargement des ontologies, fonctions d'agrégation (père) et coloration des concepts.
  - Visualisation des graphes de concepts.

**⚠ Pour l'heure, les dimensions Stop, Move, Lieux et Accompagnements sont fixées.**

Une nouvelle phase de développement doit être mise en place pour permettre la détection automatique de dimension.

Une seconde amélioration serait une meilleure gestion des dimensions d'agrégation qui sont, pour l'heure, calculées en temps réel (méthode `agg_id`).

Un stockage par dictionnaire ou directement depuis le fichier de l'ontologie du noeud d'agrégation des concepts permettrait de faciliter les manipulations et améliorer la rapidité d'exécution.

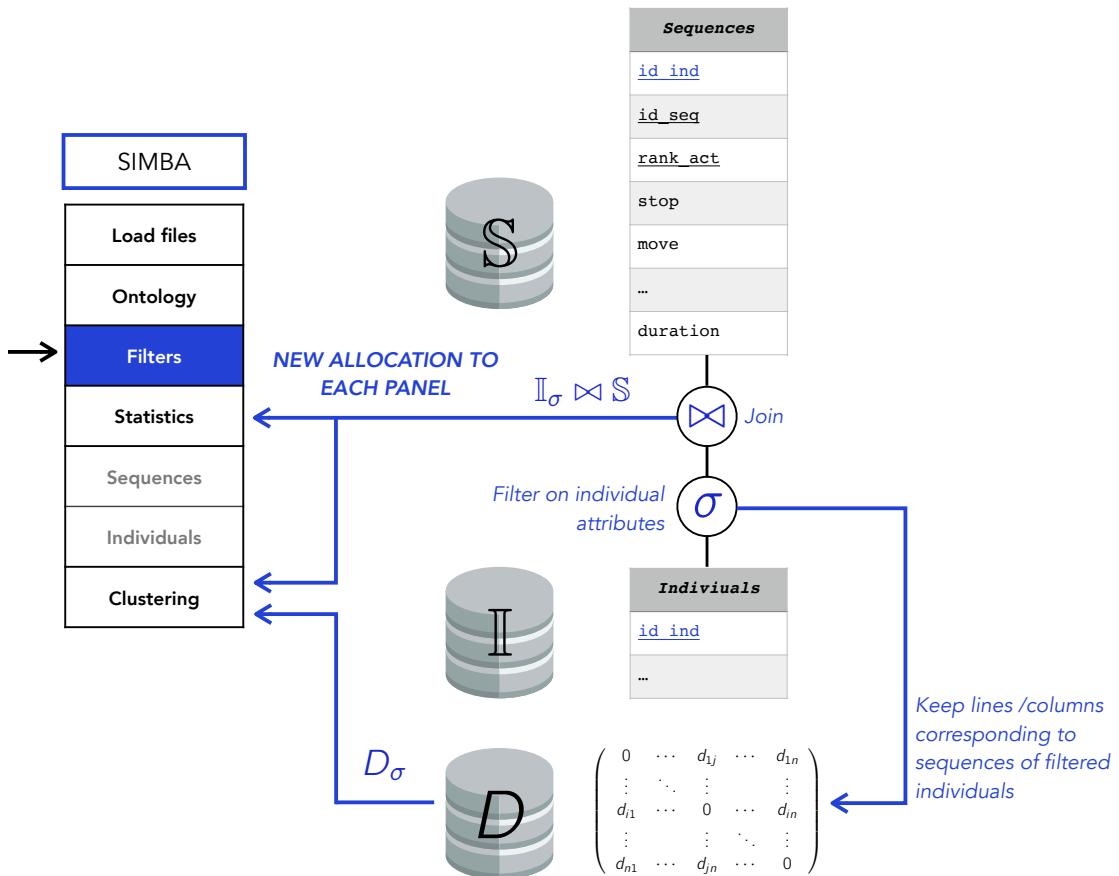
## 3. Filtrage de données

La troisième section de SIMBA, dédiée au filtrage des données, permet à l'utilisateur l'ajout de conditions logiques afin de sélectionner les individus et séquences obéissant à l'ensemble des critères exigés.

La figure ci-dessous représente le processus de filtrage et de mise à jour des fichiers de l'application. L'ensemble des filtres applicables, noté  $\sigma$ , concerne l'ensemble des données contextuelles socio-démographiques  $\mathbb{I}$ . On obtient ainsi une vue des données,  $\mathbb{I}_\sigma$ , combinée ensuite par jointure à l'ensemble des séquences  $\mathbb{S}$  tel que  $\mathbb{I}_\sigma \bowtie \mathbb{S}$ . Cet ensemble est ensuite alloué aux sections *Statistics* et *Clustering* pour le traitement et l'analyse de données. Parallèlement, la matrice de distance  $D$  est également filtrée afin de conserver uniquement les couples de séquences issus des individus de  $\mathbb{I}_\sigma$ . On obtient une matrice  $D_\sigma$  ré-allouée à la section *Clustering*.

Un point important est que, pour l'heure, le filtrage conditionnel implémenté **n'autorise que la condition d'égalité**. Un corolaire de cette contrainte est que le filtrage sur les valeurs numériques est impraticable en pratique. Une nouvelle phase de développement doit être mis en place pour palier ce problème.

La figure ci-dessous montre l'interface de la section Filters de SIMBA. On voit ici plusieurs filtres appliquées sur les données MOBI'KIDS où l'utilisateur a conservé les individus enfants (`pers_parente = Un enfant`) et de sexe féminin (`pers_sexe = Femme`). La partie droite de l'interface fournit une vue sur les données filtrées.



### Filtrage des données dans SIMBA

The screenshot shows the SIMBA App interface. The left sidebar includes options for Chargement des fichiers, Ontologie, Filtrer les données (selected), Statistiques Globales, and Clustering. The main area features a "Filtrer les données" (Filter data) panel with dropdown menus for "pers\_parente" (Un enfant), "pers\_sexe" (Femme selected), and a checkbox for "pers\_sexe" (Female checked). Buttons for "Ajouter filtre" (Add filter) and "Afficher/Mettre à jour" (Display/Update) are present. To the right is a data grid table with columns: fam\_id, pers\_parente, pers\_sexe, enq\_ville, enq\_ecole, and enq\_classes. The table displays 10 entries out of 3,865, with a "Show 10 entries" button. At the bottom are navigation buttons for Previous, Next, and a "Exporter les données filtrées (CSV)" (Export filtered data as CSV) button.

Interface de filtrage de SIMBA

### Point technique

L'interface et les fonctionnalités de filtrage sont assurées par les fichiers suivants :

- **simba\_ui\_filter.R**  
Design de la page
- **simba\_server\_button.R**  
Ajout / suppression de filtres conditionnels
- **simba\_server.R**  
Observer de modifications dans les filtres.

⚠ Une nouvelle phase de développement doit être mise en place pour permettre le filtrage selon les conditions logiques plus élaborées (i.e., <, >, !=).

## 4. Statistiques globales

La section *Statistiques globales* est décomposée en deux sous-parties : (i) Un sous-onglet *Sequences* reprend sous un format interactif une partie des indicateurs développés durant la thèse, entre autre :

- La **distribution du nombre d'activités journalières** (i.e., longueur de séquences).
- La **distribution globale des activités** par catégorie (e.g., stop, move, ...).
- Le diagramme de flots (**chord diagram**).
- Les motifs topologiques de déplacement récurrents (**daily pattern**).
- Les **tapis de séquences** fournissant une représentation chronologique des activités.

Le sous-onglet (ii) *Individus* permet une analyse bi-variée des variables issues de la table *Individuals*, I. L'onglet reprend les indicateurs :

- De **table de contingence** entre variables.
- De **diagramme mosaïque** (avec **résidus de Pearson**).

### Point technique

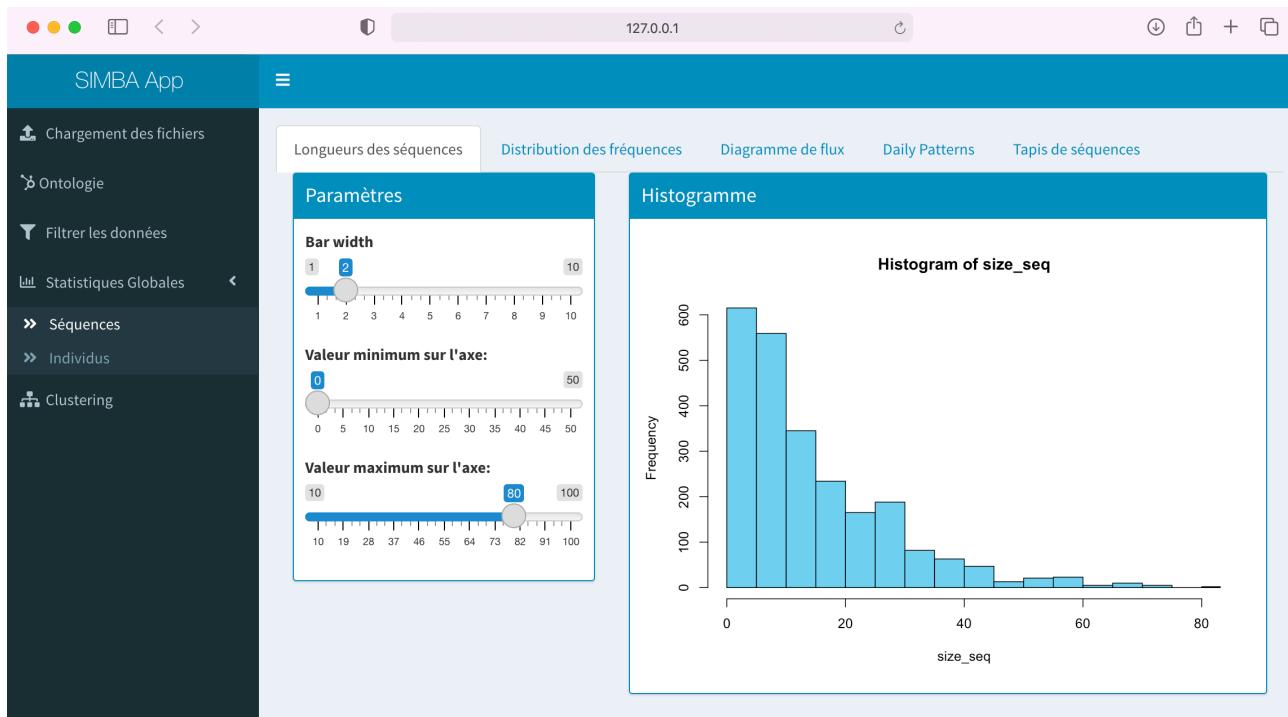
L'interface et les fonctionnalités statistiques sont assurées par les fichiers suivants :

- **simba\_ui\_stats.R**  
Design de la page.
- **simba\_server\_statsSeq.R**  
Ensemble des indicateurs sur les séquences.
- **simba\_server\_statsInd.R**  
Ensemble des indicateurs sur les individus.

## 4.1 Statistiques sur les séquences

### 4.1.1 Distribution du nombre d'activités journalières

La distribution du nombre d'activités journalières (i.e., longueur de séquences) renseigne le nombre d'activités (stop + move) effectuées au cours de la journée.



Distribution du nombre d'activités journalières

### Interprétation

Le graphique montre une distribution de la taille des séquences majoritairement concentrée sur des valeurs faibles (1 à 5). Ces individus font majoritairement peu d'activités au cours de la journée.

Ce résultat diffère légèrement du graphique obtenu dans l'étude des EMD et de l'état de l'art où la distribution du nombre d'activités journalières suit une loi de Poisson. [Rhee et al. \(2011\)](#)

### 4.1.2 Distribution globale des activités

La distribution globale des activités référence la totalité des activités au sein de chaque séquence selon la dimension sélectionnée.

Cet onglet est composé de deux graphiques :

- Le graphique de gauche présente un *barplot* à échelle du *logarithmique*. La couleur des activités fait référence à celle renseignée au sein de l'ontologie.
- Le graphique de droite montre l'adéquation à un modèle *zipfien*. Si l'ensemble des points sont alignés selon la droite, alors la distribution des données suit une loi de Zipf.



Distribution globale des activités Stop : (a) À gauche, barplot (b) À droite, ajustement à un modèle zipfien

### Interprétation

La loi suivie par la distribution des activités de STOP suit globalement le modèle zipfien établi par [Song et al. \(2010\)](#).

Toutefois, on voit la distribution s'effondrer sur les activités de fréquence les plus faible ce qui suggère une anomalie dans les relevés effectués.

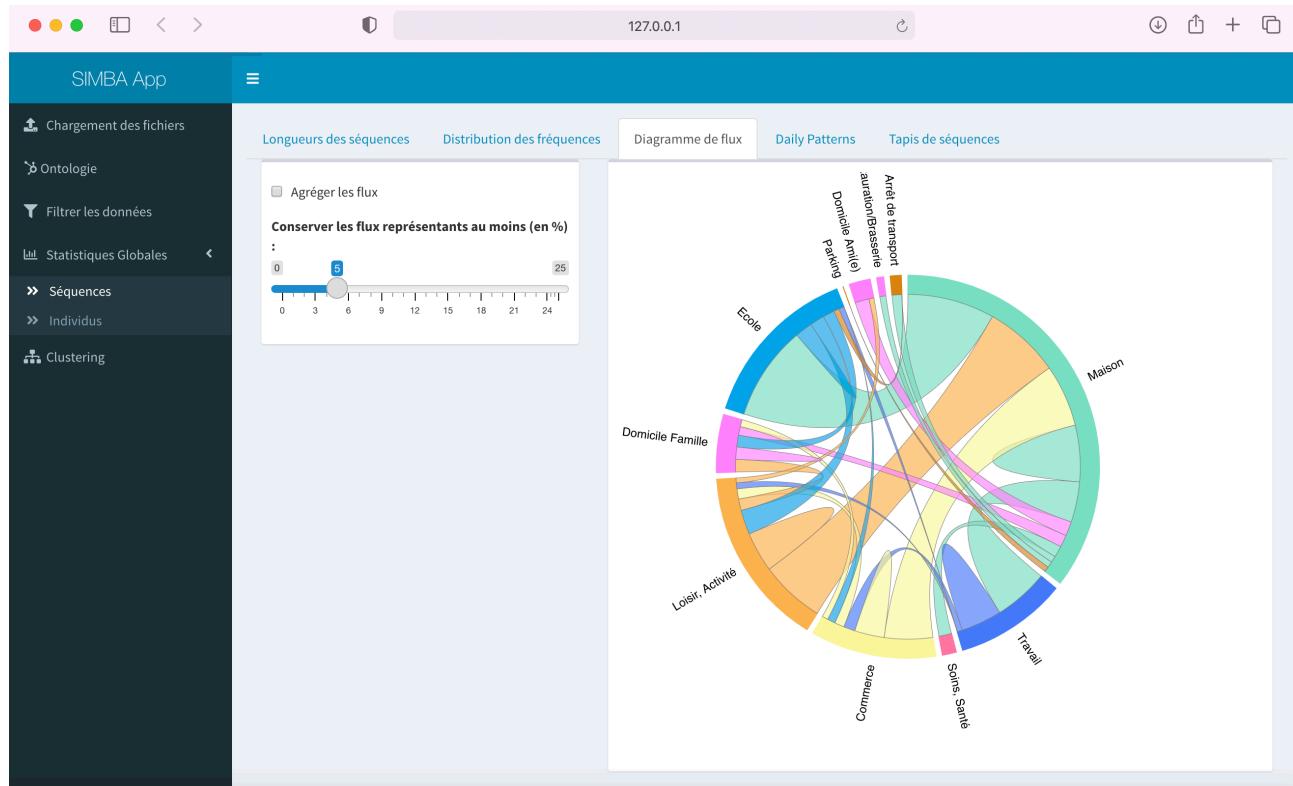
#### 4.1.3 Diagramme de flots (ou Chord Diagram)

Le diagramme de flots permet de représenter les transitions entre activités. Ici nous représentons les transitions entre deux STOP consécutifs.

Le diagramme est uniquement disponible pour la dimension **Place** car il réclame une colonne de la base de données au contenu atomique.

Survoler le flot permet de mettre en évidence le nombre de transitions de  $x \rightarrow y$  et réciproquement ( $de x \leftarrow y$ ).

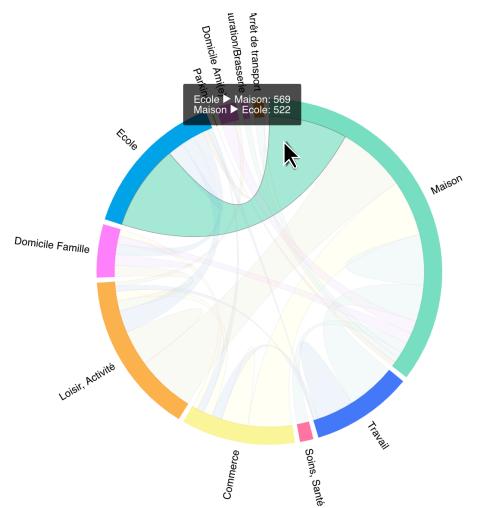
Il est possible de filtrer les flux minoritaire du graphique à l'aide d'un curseur. La valeur 0% permet d'afficher l'ensemble des transitions observées dans les données.



### Transitions entre les différents lieux

Formellement, soit la matrice Origine-Destination  $OD = \{t_{ij}\}$  où  $t_{ij}$  est le nombre de transitions, le pourcentage de conservation  $\rho$  permet de d'afficher les flux tels que:

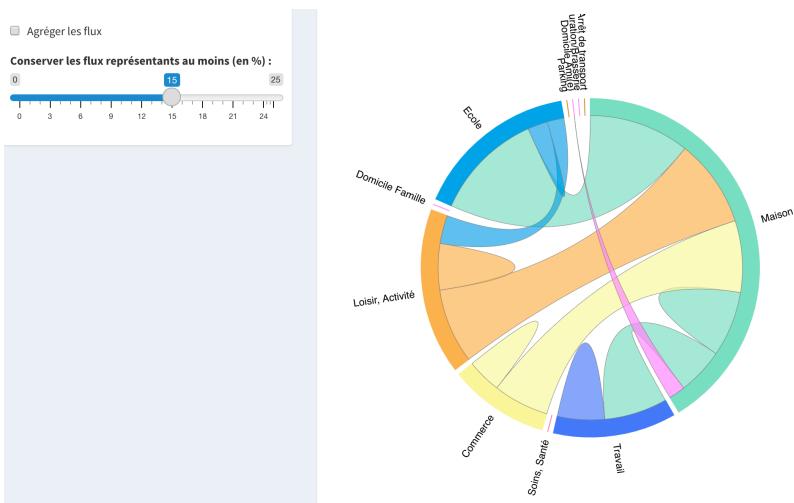
$$t_{ij} \geq \frac{\rho \times \max_{i,j}\{t_{ij}\}}{100}$$



Mouse hover un flux.

On constate que 569 individus passe de l'école à la maison

Réduction des flux à  $\rho = 15\%$  du flux maximal



### Interprétation

On constate une répartition des transitions assez semblable à celle relevée dans le jeu de l'EMD 2018.

On notera toutefois la présence d'auto-flots au niveau de la catégorie « domicile », peu observés sur le jeu de l'EMD. Ainsi qu'un nombre de transitions importants entre les catégories « domicile » et « Loisir, activité ».

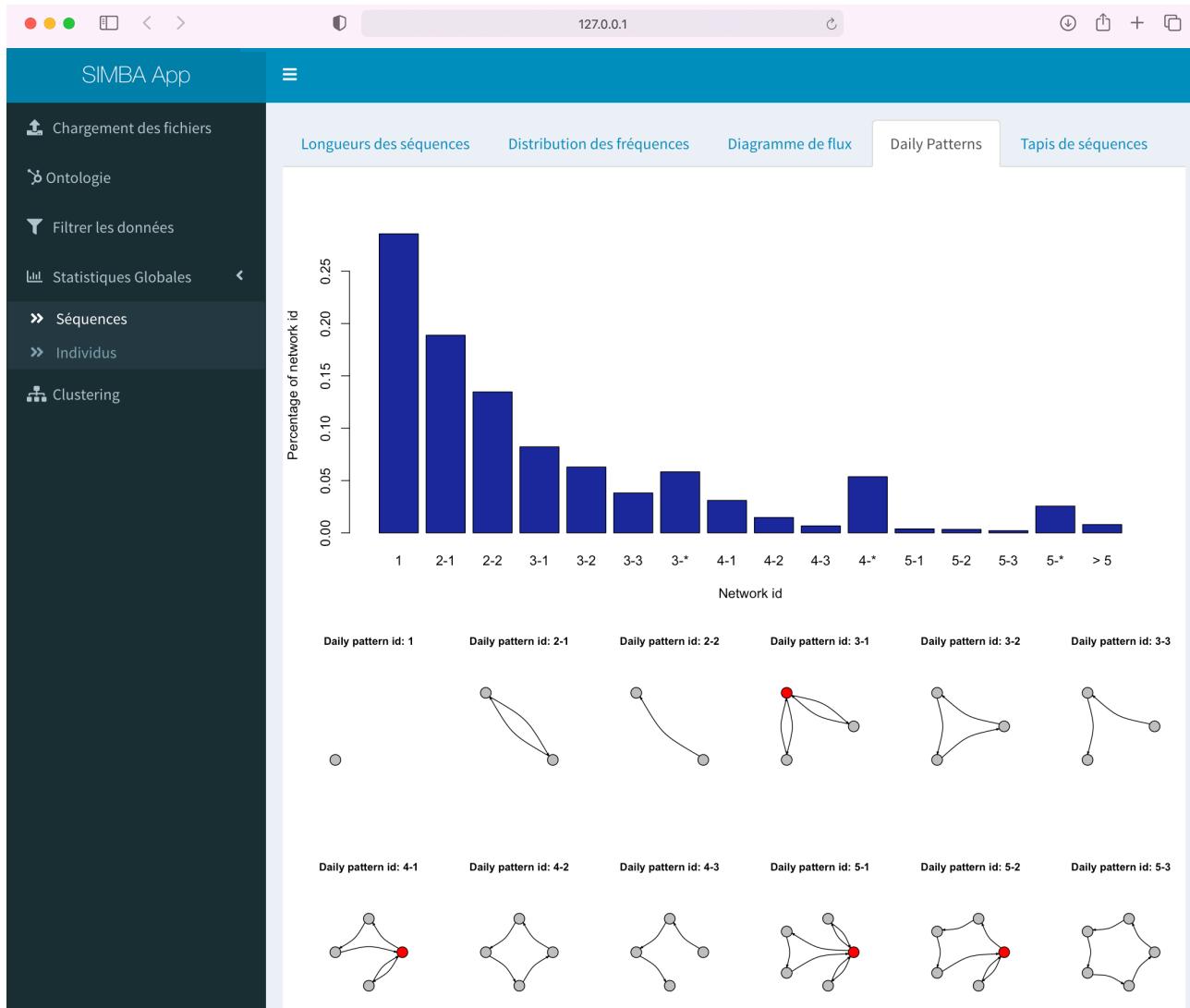
#### 4.1.4 Motifs topologique (ou Daily pattern)

Les motifs topologiques sont un concept établi par Schneider et al. (2013) et viennent renseigner la forme des déplacements quotidiens, au niveau des STOP: s'agit-il d'une oscillation entre deux lieux ? Trois lieux consécutifs visités ? etc.

Un tel indicateur est capable à la fois de capter à la fois la diversité des déplacements (nombre de lieux différents visités), leur forme mais aussi la cyclicité potentielle des déplacements (i.e., existe-t-il des oscillations / boucle entre deux ou plusieurs lieux).

Tout comme le diagramme de flot, les daily patterns renseignent uniquement la dimension **Place** car ils ne peuvent être calculés uniquement avec des données au contenu atomique.

Par exemple, dans la figure ci-dessous, celle-ci montre qu'environ 20% des déplacements sont formés d'oscillation unique de la forme → → → → où l'on constate un aller-retour entre le domicile et l'école.



Distribution des daily patterns référencés dans les séquences journalières Mobi'Kids

## Interprétation

Le graphique montre une distribution des motifs très différente à la fois des résultats donnés par Schneider et al. mais aussi de celle trouvée dans l'EMD.

En particulier le motif 1 (i.e., rester chez soi) est très nettement représenté à hauteur ~30% ce qui est en accord avec les résultats de Schneider. Rappelons que ce motif n'est pas représenté dans l'EMD.

Cependant, la fréquence des autres motifs est anormal au regard des résultats des études paires. Notamment, les boucles oscillantes (e.g., motifs 2-1, 3-1) sont nettement sous-représentés.

Également, des motifs assez surprenants non cycliques (e.g., motifs 2-2, 3-3) sont, à l'inverse, sur-représentés.

### Point technique

⚠ Le graphique des Daily Patterns, pour l'heure, ne peut se mettre à jour selon le filtrage des données effectué par l'utilisateur.

La raison de ce fait est que le calcul des Daily Patterns est subordonné à un script Python qui prend en entrée le fichier initial des séquences sémantiques journalières.

Une piste simple pour pallier ce problème serait d'exporter un fichier .csv des données filtrés par l'utilisateur dans le répertoire Data puis de tester l'existence d'un tel fichier dans le script Python de calcul des Daily Pattern. Si ce fichier existe, il est pris en priorité pour effectuer les calculs.

#### 4.1.5 Tapis de séquences

Les tapis de séquences sont une méthode de visualisation et d'analyse de données séquentielles implémentée dans le package R TraMineR.

Ce mode de visualisation n'est **pas encore implémenté** dans SIMBA.

Toutefois, nous évoquons quelques pistes pour aider au développement de cet indicateur, les difficultés techniques ainsi que quelques briques de développement préliminaires afin de faciliter son intégration.

### Point technique

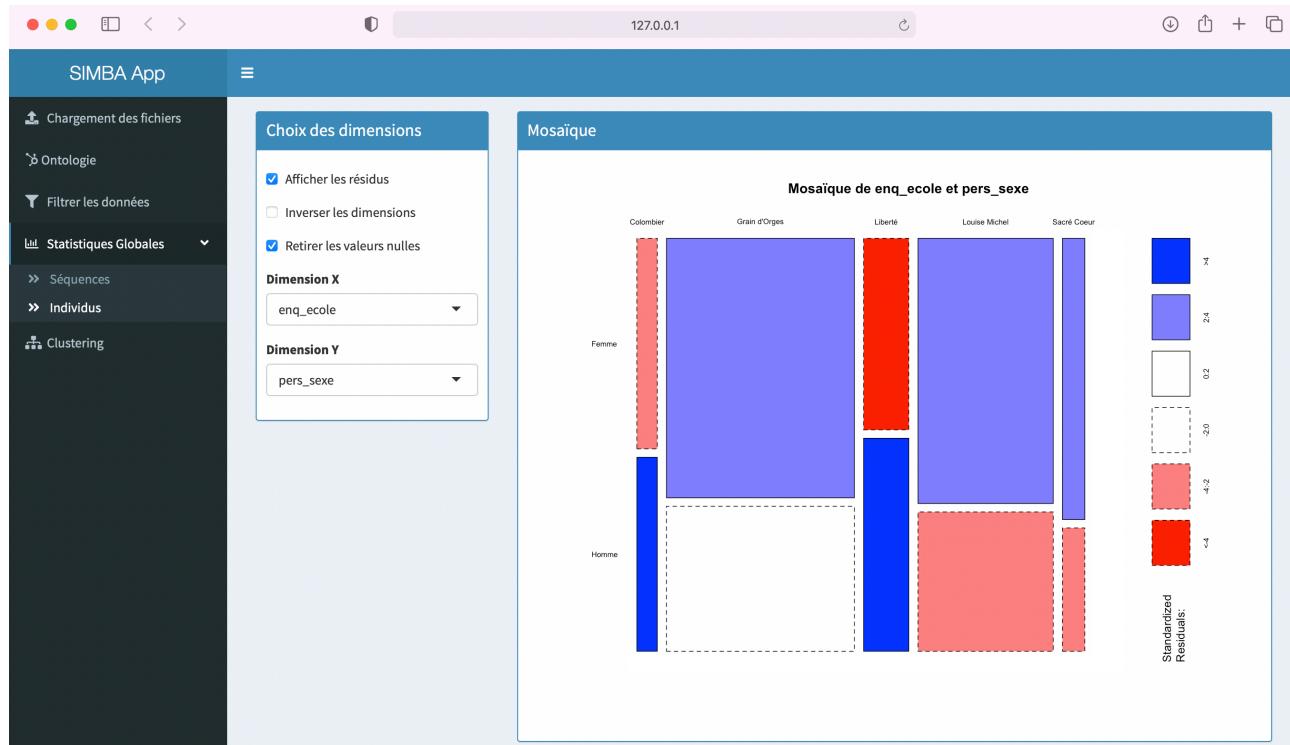
Nous évoquons d'abord les difficultés techniques de cette représentation.

1. La granularité temporelle définie dans le jeu Mobi'kids (1 min). De fait, le format d'entrée des séquences au sein du TraMineR réclame 1 colonne par unité de temps. Nous avons alors ici 1440 colonnes (1 colonne / min sur toute une journée).  
Or, la bibliothèque ne permet pas de représenter des données aussi larges. Deux solutions sont possibles pour résoudre ce problème:
  - i. Adopter une unité atomique de temps plus grande (e.g., 5min comme dans l'EMD).
  - ii. Représenter le tapis de séquences uniquement sur un intervalle de temps fixe (e.g., 4h) où le début / fin seraient réglables par l'utilisateur à l'aide de sliders (voir figure section 4.1.1).
3. La représentation des séquences sous forme d'ensemble de concepts. De fait, nous devons encore ici restreindre notre analyse au dimension atomique **Place**. Concernant les modes de déplacement, ceux-ci doivent être ensuite agrégé pour conserver un unique label afin d'évoluer sur des séquences standards type STOP-MOVE.
4. Les nombreuses activités manquantes créent des « trous » dans les séquences. Une méthode d'imputation des données doit être mis en place pour combler ces valeurs manquantes (voir rapport technique Qualité et préparation des données)

De premières pistes de développement peuvent être trouvées sur le Github de SIMBA, notamment le fichier `toTraminer.py` qui permet de créer un fichier .csv de l'ensemble des données Mobi'Kids sous un format assimilable par le package TraMineR.

## 4.2 Statistiques sur les individus

L'onglet dédié aux statistiques sur les individus permet de réaliser une analyse bi-variée par diagramme mosaïque avec test de significativité (i.e., résidus de Pearson) afin de détecter les liens statistiques entre variables et modalités issues du fichier I des individus.



### Analyse bi-variée du sexe des individus interrogés et de leur école d'origine

Cette fonctionnalité a la double utilité de permettre de détecter les potentiels biais statistiques préalables dans les données grâce aux résidus de Pearson. Par exemple, la figure ci-dessus montre que plus d'enfants garçons ont été interrogés à l'école « Liberté », mais aussi d'établir les proportions entre les différentes modalités et variables par l'usage du diagramme mosaïque.

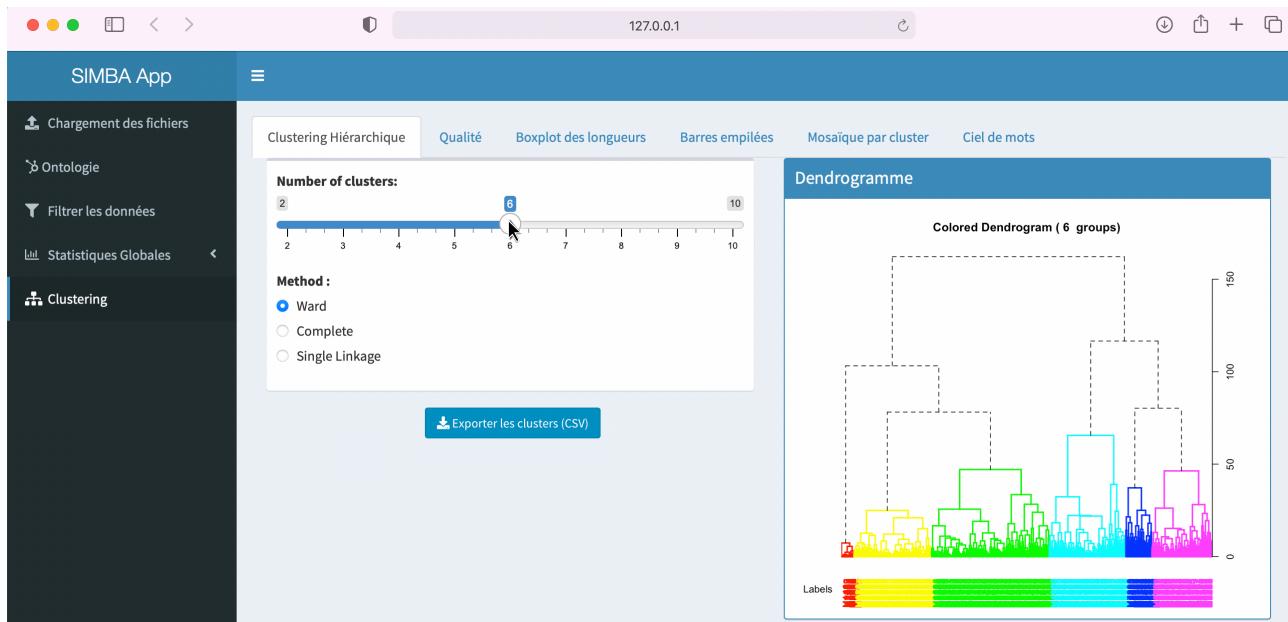
### Point technique

Les points suivants détaillent les améliorations futures pouvant être accomplies au sein de cette section :

- Une refonte du code du fichier `simba_server_statsInd.R` serait souhaitable, notamment par l'usage de requête ajax et séparer clairement les fonctionnalités entre client et server.
- L'ajout d'une table de contingence ou d'un diagramme mosaïque dynamique (à l'aide de mouse hover) permettant de renseigner quantitativement les effectifs serait le bienvenu en complément du diagramme mosaïque actuel.
- L'ajout de nouveaux indicateurs pour effectuer des analyses univariées serait souhaitable.

## 5. Clustering

La dernière section concentre les outils liés au clustering et à la fouille interactive de comportements issus des séquences sémantiques. Pour l'heure, SIMBA propose la réalisation d'un clustering hiérarchique selon les trois critères principaux d'agrégation (Ward, Single, Complete). Le nombre de clusters est géré dynamiquement par l'utilisateur. Du processus de clustering résulte une table `Cluster[id_seq, id_clust]` (notée C) qui associe à une séquence un numéro de cluster. Enfin, un bouton « export » permet le téléchargement d'un fichier .csv de la forme  $(\mathbb{I}_{\sigma} \bowtie \mathbb{S}) \bowtie \mathbb{C}$ , soit l'ensemble des séquences filtrées + informations sur les individus filtrés + numéro de cluster de la séquence, afin de poursuivre l'analyse de façon indépendante.



Page d'accueil de la section clustering — Clustering hiérarchique

La figure ci-dessus présente la page consacrée au clustering des séquences. L'utilisateur choisit le nombre de clusters, le critère d'agrégation et dispose de l'affichage en direct du dendrogramme et des clusters formés. En tant qu'algorithme dont le résultat est visualisable, nous pensons que le clustering hiérarchique est la méthode la plus appréhendable par l'utilisateur pour notre tâche de regroupement de séquences sémantiques, c'est pourquoi elle est disponible en priorité dans SIMBA. La section propose 4 autres onglets pour l'analyse des clusters :

- L'analyse de la *qualité des clusters*.
- L'analyse des activités par cluster et par dimension à l'aide de diagramme à *barres empilées*.
- L'analyse bivariée par *diagramme mosaïque* entre les clusters et une variable qualitative sélectionnée par l'utilisateur.
- Un résumé des clusters par *ciel de mots*.

## Point technique

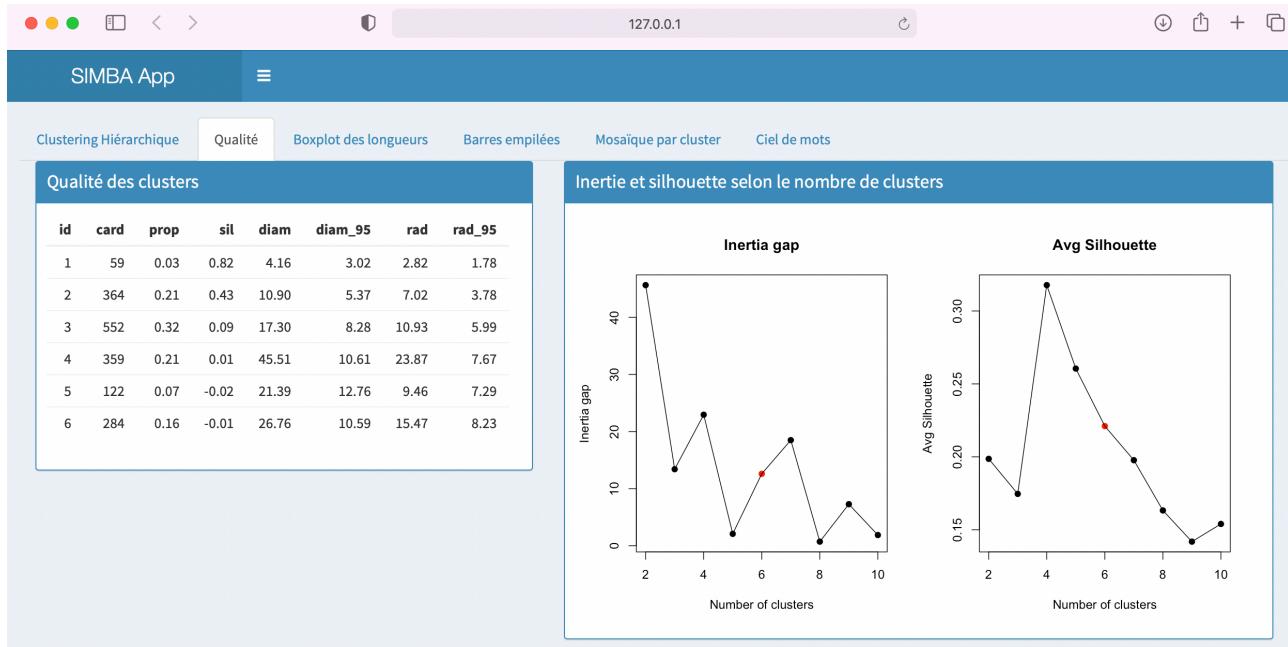
Les fonctionnalités de la section clustering sont assurés par les fichiers :

- **simba\_ui\_cluster.R**  
Design et rendu visuel de la page
- **simba\_server\_cluster.R**  
Différentes fonctionnalités et onglets de la section.

### 5.1 Qualité des clusters

L'onglet de qualité des clusters permet de visualiser quelques informations sur la forme et la densité des clusters. Il constitué de deux panneaux principaux :

- Une table dotée des informations :
  - L'id de cluster
  - La cardinalité du cluster (i.e., le nombre de séquences qu'il contient).
  - La proportion du cluster (i.e., sa cardinalité sur le nombre total de séquences).
  - L'indicateur Silhouette du cluster.
  - Le diamètre du cluster (i.e., la distance maximale entre les deux éléments les plus éloignés du cluster).
  - Le diamètre 95% (i.e., Idem mais au 95ème percentile de la distribution des distances).
  - Le rayon du cluster (i.e., la distance maximale entre le medoid du cluster et l'élément le plus éloigné de celui-ci).
  - Le rayon 95% (Idem mais au 95ème percentile de la distribution des distances).
- Les graphiques de saut d'inertie et de silhouette moyen. On cherche à maximiser ces quantités. Le point rouge représente le nombre de clusters actuellement sélectionné par l'utilisateur.



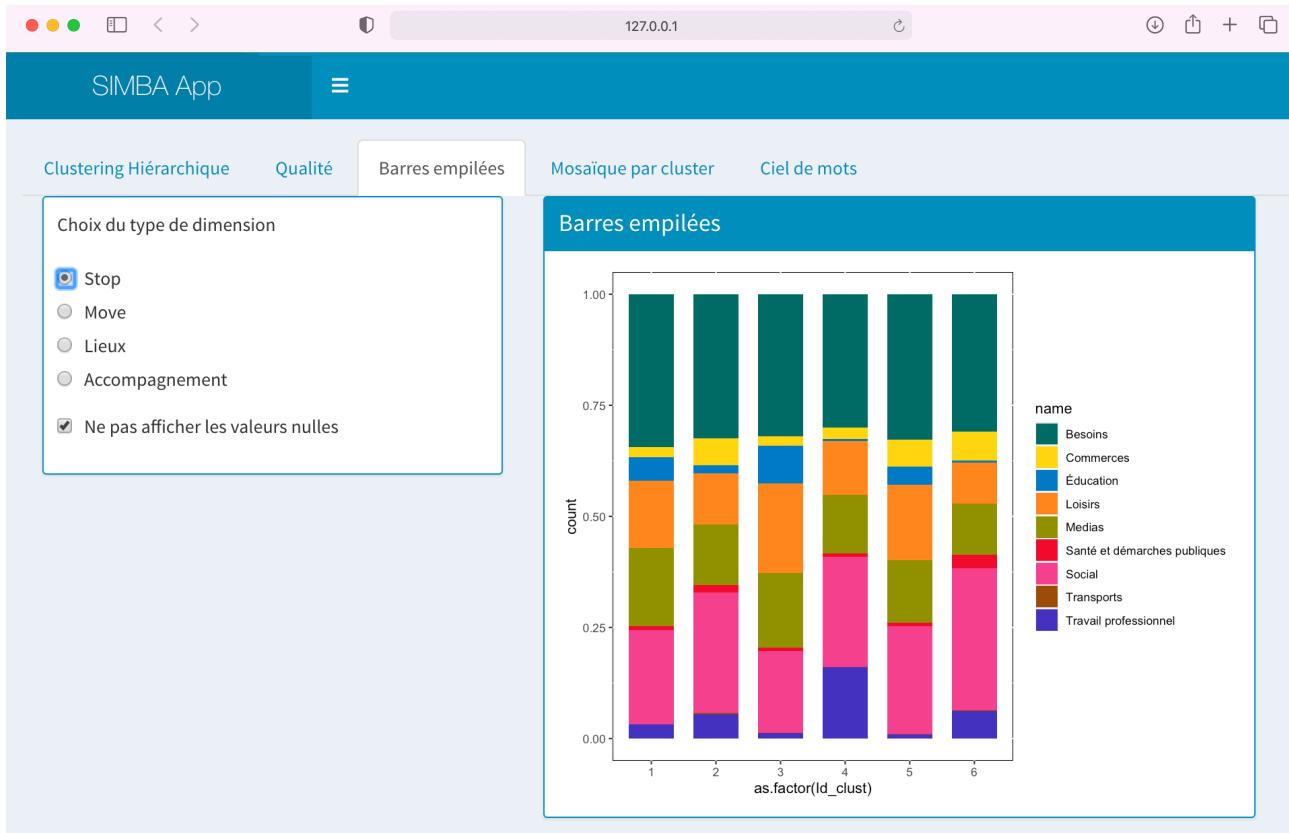
Onglet de qualité de cluster (6 clusters sélectionnés)

La quantité Silhouette mesure un rapport entre la compacité du cluster et la séparation des autres clusters. Plus cet indicateur tend vers 1, plus celui-ci est bon. À l'inverse, plus celui-ci tend vers -1, plus le cluster est mal formé selon les critères de Silhouette (i.e., compacité et séparation).

## 5.2 Barres empilées et analyse d'activités par cluster

Les barres empilées permettent de visualiser les proportions du nombre d'activités effectuées au sein de chaque séquence par cluster.

Par exemple, le graphique ci-dessous montre que le cluster 3 contient davantage d'activités labellisées « Éducation » que les autres clusters.



Barres empilées des activités STOP pour les 6 clusters extraits

### Point technique

Les points suivants détaillent les améliorations futures pouvant être accomplies pour ce graphique :

- Le temps de calcul des barres empilées est anormalement long. Une optimisation des algorithmes de calcul est requis.
- Inclure l'option d'agréger ou non les activités dans l'ontologie. Par défaut ici les activités sont agrégées.
- Passer le graphique en version dynamique en utilisant la bibliothèque `ggplotly`.
- Afin d'analyser également la dimension temporelle, une amélioration notable serait d'inclure la possibilité pour l'utilisateur (via une case à cocher) de représenter non plus le nombre d'occurrence, mais les budgets temps associés à chaque activité.

### 5.3 Diagramme mosaïque par cluster

L'onglet « Mosaïque par cluster » permet d'établir un diagramme mosaïque avec résidus de Pearson d'une variable sélectionnée par l'utilisateur (sur la figure ci-dessous le niveau scolaire) avec l'ensemble des clusters constitués.

Cette analyse permet de dévoiler les liens statistiques entre les clusters et la variable sélectionnée.



Analyse par diagramme mosaïque de la répartition des niveaux scolaires dans les clusters

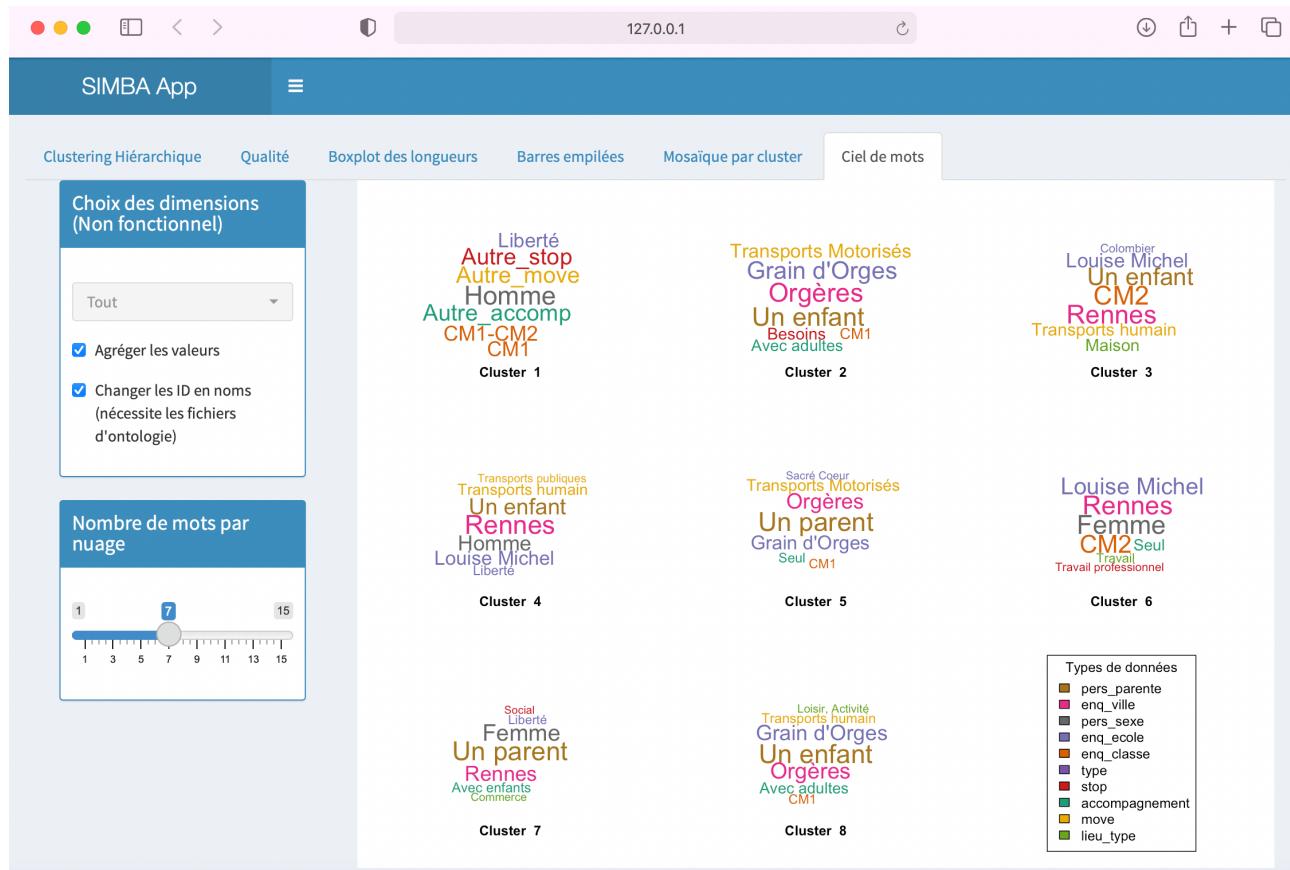
Nous renvoyons à la section 4.2 qui détaille les points techniques améliorable sur cet indicateur.

### 5.4 Nuages et ciel de mots

Le ciel de mots représente un ensemble de nuages de mots descriptifs de chacun des clusters. Ces mots sont choisis par le biais d'un score qui représente une combinaison linéaire du résidu<sup>1</sup>  $r$  de Pearson de la modalité  $m$  (i.e., mot) à l'intérieur du cluster et de sa fréquence  $f$ , soit :

$$score(m) = \begin{cases} f(m) \times r(m) & \text{Si } r(m) \geq 0 \\ f(m) \times \frac{1}{|r(m)|} & \text{Si } r(m) < 0 \end{cases}$$

<sup>1</sup> On rappelle que le résidu  $r$  de  $m$  s'exprime tel que :  $\frac{f(m) - f^*(m)}{\sqrt{f^*(m)}}$  où  $f^*(m)$  représente la fréquence théorique de  $m$ .



Ciel de mots constitué de 8 nuages représentant les clusters

Les mots sont ensuite triés par score et par dimension (stop, move, accompagnement + variables issues de la base I des individus) ; un slider permet de sélectionner le nombre de mots à afficher par cluster.

Des tests utilisateurs et études approfondies d'un point de vue cognitif et du ressenti humain sont en cours afin d'adapter les modes de visualisation et contrôler la pertinence des informations rendues à l'utilisateur et trouver un juste équilibre entre qualité de restitution, ergonomie et esthétisme.

### Point technique

Le code en charge de la fonctionnalité du ciel de mots est le fichier :

- `simba_server_ciel.R`

**⚠️** Le développement de cet indicateur souffre de nombreux écueils. Une nouvelle phase de développement serait nécessaire afin de rendre celui-ci optimal.

Une première piste serait de repartir de la définition mathématique des résidus et du score ci-dessus afin de gérer ensuite l'affichage en fonction de ce dernier.

## 6. Conclusions et perspectives

SIMBA forme pour l'heure un prototype d'application web pour l'analyse et la fouille dynamique de séquences sémantiques prometteur.

Néanmoins, certaines fonctionnalités nécessitent encore quelques ajustements et phases de développement, nous noterons en priorité les développements suivants :

- Gestion automatique des dimensions au sein des ontologies (voir section 2.2)
- Expressivité des conditions de filtrage (voir section 3)
- Ajustement des indicateurs statistiques: Daily patterns et Tapis de séquences (voir sections 4.1.4 et 4.1.5)
- Développement ajax pour le diagramme mosaïque et l'analyse bi-variée des individus (voir section 4.2)
- Amélioration des performances et refonte du code des indicateurs statistiques pour le clustering: Barres empilées, Diagramme mosaïque et Ciel de mots (voir sections 5.2, 5.3 et 5.4)

De façon générale, une **refonte du code de SIMBA** à des fins d'optimisation et de clarté serait souhaitable.

Le lecteur attentif notera également un point bloquant récurrent. Bien que SIMBA supporte la gestion de séquences multi-ensembles, ce fait pose de nombreux problèmes aux indicateurs statistiques sur les séquences qui, bien souvent, doivent se contenter d'analyser l'unique dimension atomique disponible (Place).

De fait, une **structuration des séquences sur le modèle des EMD** (dimension STOP + MOVE atomiques) permettrait de mener des analyses plus *intelligibles et complètes*.

De nombreuses perspectives de développement peuvent être évoquées. Celles-ci sont détaillées en section 8.3.3 de la thèse. Notons en autres :

- L'automatisation du calcul de la matrice de distance selon une mesure disponible dans le package TraMineR.
- L'intégration d'indicateurs temporels (tapis de séquences, budgets-temps, etc.)
- Visualisation et calcul de séquences prototypiques.
- Résumé automatique de comportements

Étant donné les analyses futures sur les distances FTH et CED devant être menées. La priorité de développement devrait être portée sur l'**intégration d'indicateurs temporels** comme la visualisation par tapis de séquences.