Similarité des programmes d'activités

5e colloque annuel de la Chaire Mobilité – 26 mai 2015

Farhana Yasmin
Candidate au doctorat

Superviseurs
Prof. Catherine Morency, Polytechnique Montréal
Prof. Matthew J. Roorda, University of Toronto

mobilité

Plan

- Contexte et motivations
- Question de recherche
- Données et méthodologie
- Résultats (expérimentation)
- Contributions potentielles

Contexte et motivations (1)

- Objectif principal de la recherche → contribuer à:
  - l'amélioration de l'approche actuelle de modélisation de la demande (suites d'un projet de recherche avec le MTQ);
  - l'amélioration du cadre de modélisation selon une approche basée activités.
- Transférer **TASHA** (Travel Activity Scheduler for Household Agents – Toronto model) au contexte de l'Île de Montréal.
- **TASHA** – bonne transférabilité dans le cas des **activités contraintes** (travail, études);

  – MAIS, **moins convaincant** pour le reproduction des activités **non-contraintes** (magasinage, autre).

Contexte et motivations (2)

- Dans TASHA, la génération des activités est basée sur des **distributions observées d'activités** (fréquence, heure de départ et durée) tirées des enquêtes Origine-Destination→ Les segments de population utilisés pour construire des distributions pour chaque type d'activité **sont déterminés intuitivement**→ les segments **sont constants à travers le temps** → les **distributions** sont considérées **constantes** à travers le temps.
- MAIS, une analyse des tendances (1998 à 2008) – a confirmé que les **distributions d'activités changent** à travers le temps à Montréal.

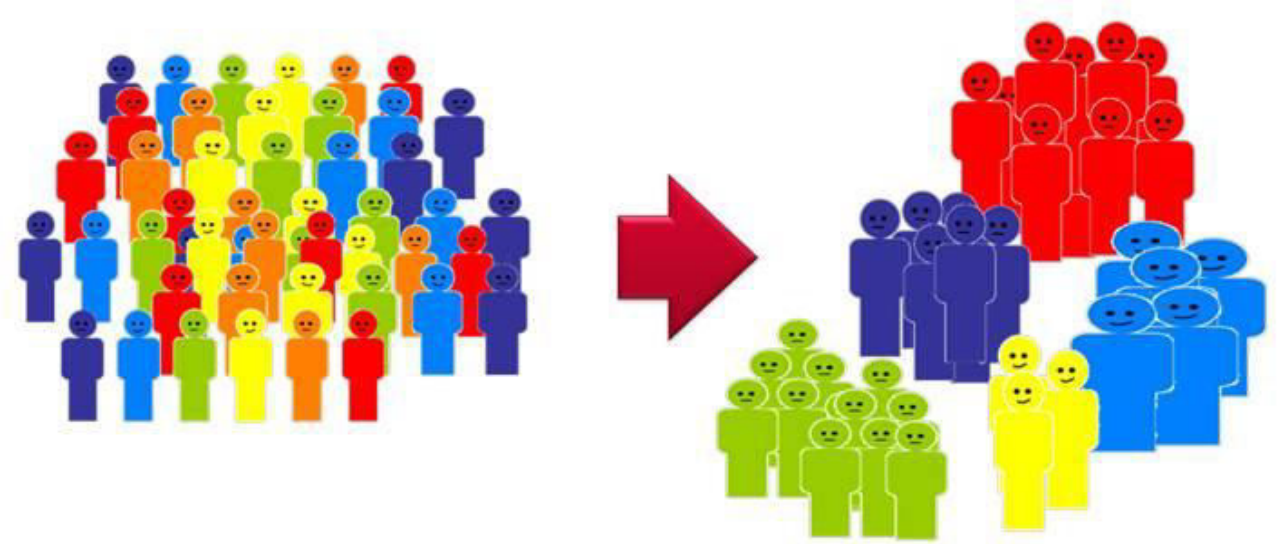
Contexte et motivation (3)

- Nécessaire d'analyser quels segments de population on des patrons d'activités similaires à travers le temps.
- Utile d'utiliser des techniques de segmentation (clustering) – bon outil pour mieux comprendre la **similarité des choix et comportements des individus.**
- L'utilisation des processus de segmentation dans les modèles d'activités est **encore rare.**

Question de recherche

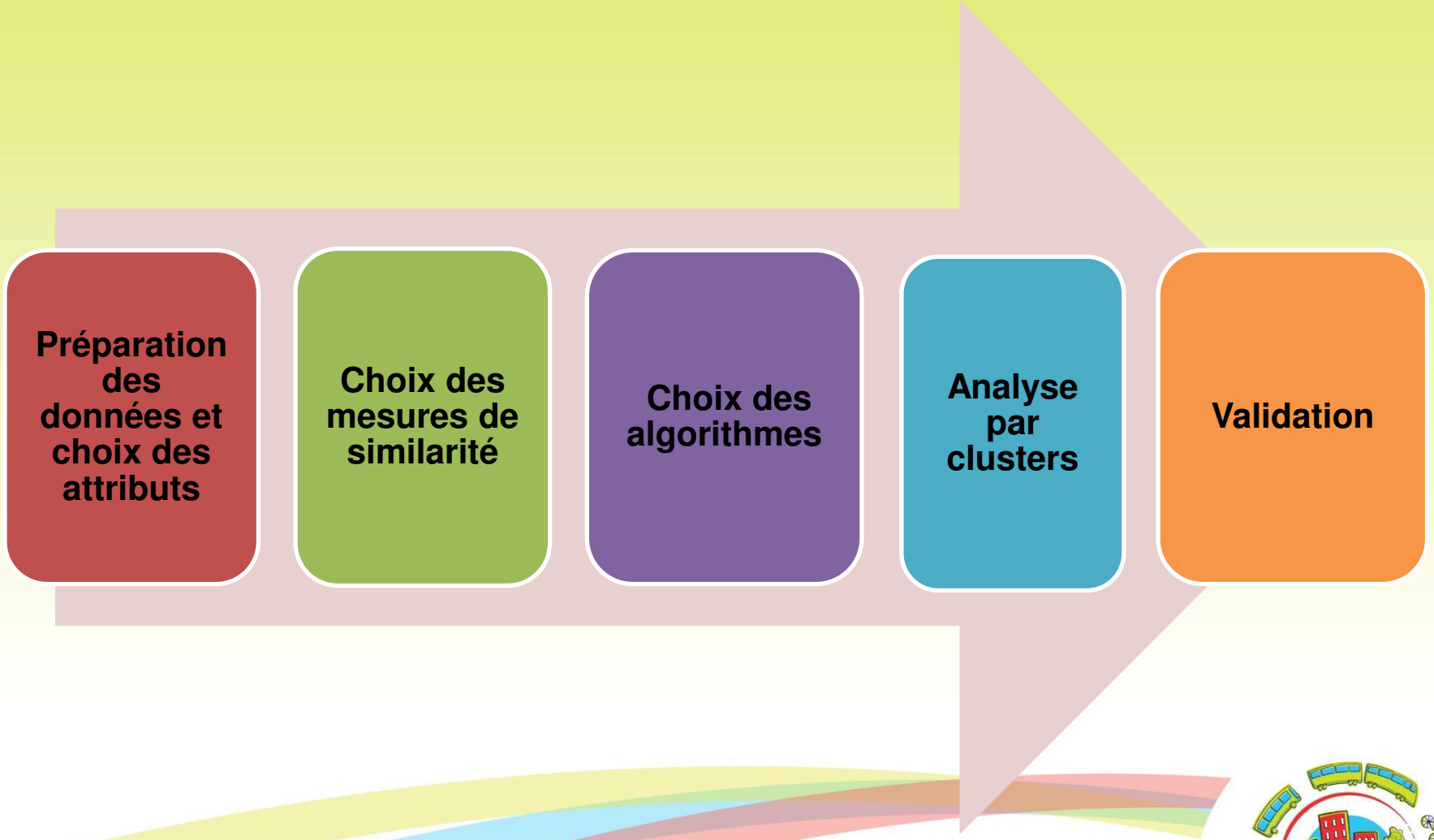
- Existe-t-il une approche optimale de segmentation de la population basée sur la similarité des patrons individuels d'activités quotidiennes pour développer les distributions requises pour la modélisation?

Fréquence des activités, heure de départ et durée des activités



Source: https://socialfs.wordpress.com/

Données et méthodologie (1)



Données et méthodologie (2)

1. Préparation des données et choix des attributs

- Étude de cas – **Grande région de Montréal**
- Échantillonnage - Sélection aléatoire de **10 000 ménages** de l'enquête OD 2003
- Variables (personnes) – **fréquence d'activités, heure de début et durée**
- **Normalisation et pondération**

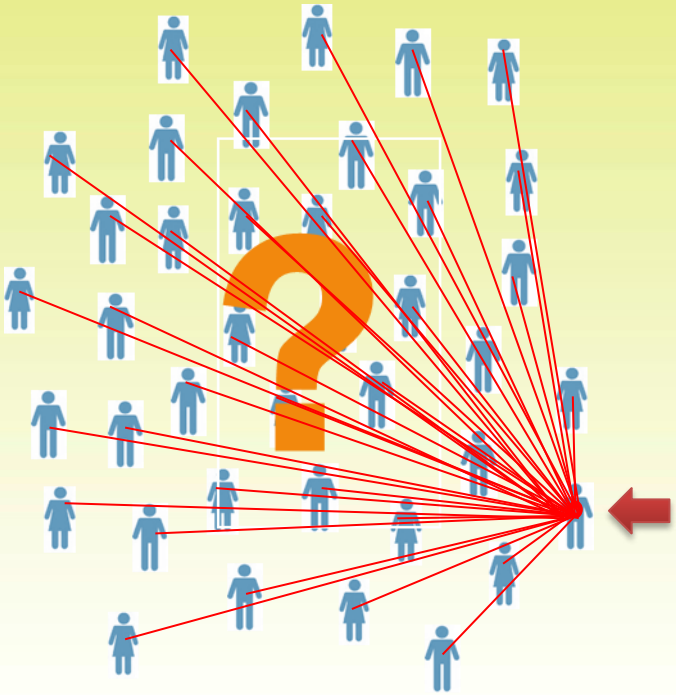
$$Normalized\ a_i' = \frac{a_i - A_{min}}{A_{max} - A_{min}}$$

Variables (unit)	Value ranges
Frequency (number)	0 - 12
Start time (hhmm)	0 - 2400
Duration (minutes)	0 - 1440

où,
a_i = valeur originale
a_i' = valeur normalisée
A_{max} = valeur maximale de la variable A
A_{min} = valeur minimale de la variable A

Données et méthodologie (3)

2. Mesures de similarité (différences)

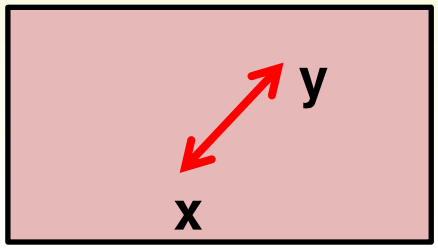


- Fonction de distance euclidienne (vol d'oiseau)

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Distance euclidienne au carré

$$D'_{Euclidean} = \sum_{i=1}^{n}(x_i - y_i)^2$$



Distance à vol d'oiseau

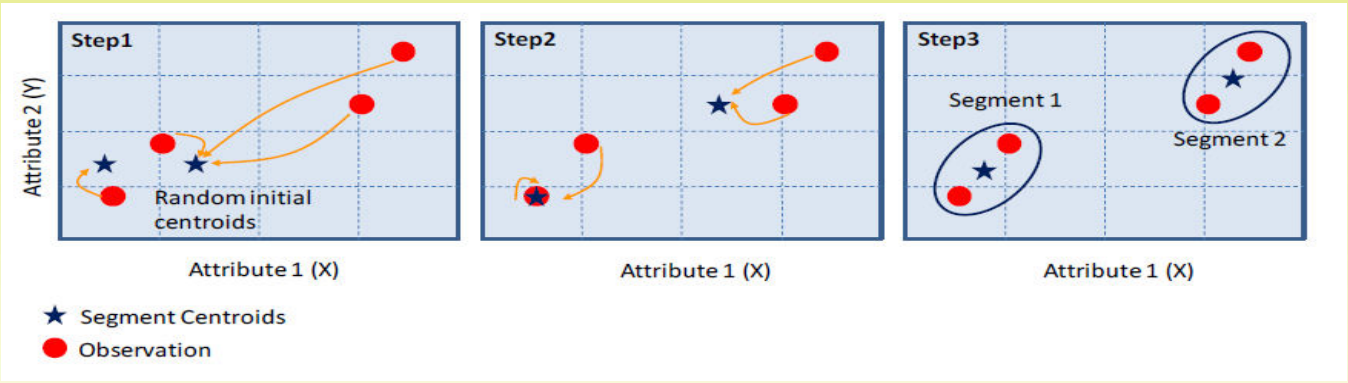
- créer une matrice de distances contenant les distances entre toutes les paires d'individus selon le vecteur d'attributs choisi.

Source: http://fbppc.com/facebook/use-unpublished-page-posts-to-sharpen-segmentation-and-drive-roi/

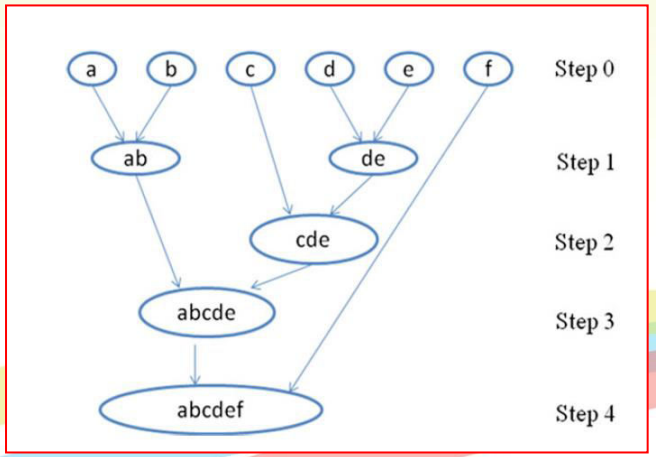
Données et méthodologie (4)

3. Algorithme et analyse par clusters

- **Méthode de partitionnement – <u>Algorithme K-moyen</u>**



- **Méthode hiérarchique par agglomération**



- **Méthode de groupement par par non pondérée avec moyenne arithmétique <u>(UPGMA)</u>**
- **<u>Méthode de Ward</u>**

Données et méthodologie (5)

4. Validation – pour évaluer la qualité du groupement

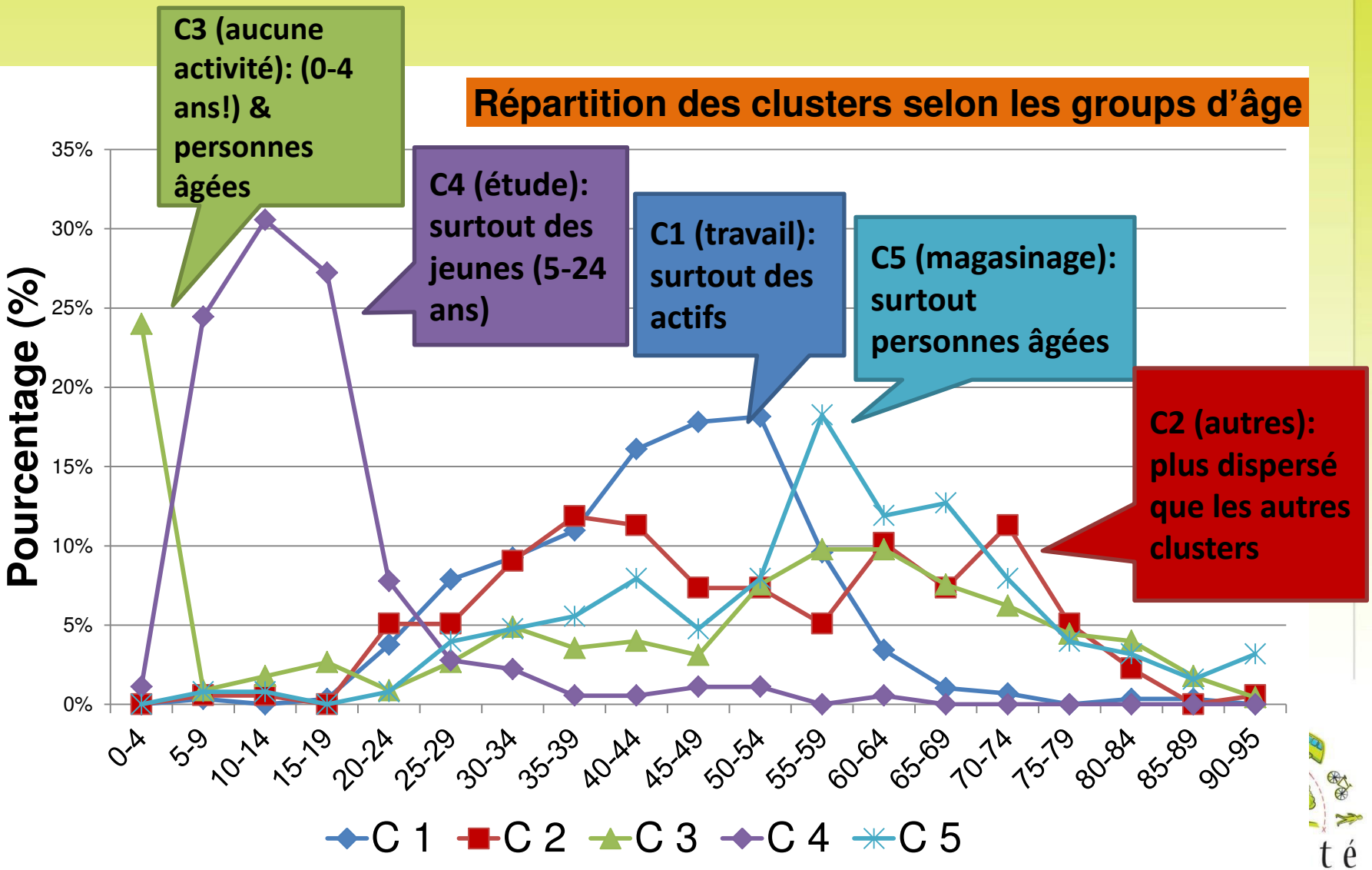
- **Mesures externes:**
  - **Statistique de Rand corrigée**
  - **Index de variation de l'information**
- **Mesures internes:**
  - **Mesure de silhouette**

Résultats du clustering: Test (1)

Échantillon: **1000 personnes (OD 2003)** Variable: **fréquence d'activités**
Méthode: **Ward** **Mesure de similarité: distance euclidienne au carré**

Cluster no.	Travail (%)	Étude (%)	Magasinage (%)	Autre (%)	Personnes (%)	Description des clusters
C1	**100%**	0%	1%	3%	**292 (29.2%)**	Principalement des activités TRAVAIL
C2	36%	0%	5%	**100%**	**177 (17.7%)**	Principalement des activités AUTRES
C3	0%	0%	0%	0%	**225 (22.5%)**	AUCUNE activité
C4	4%	**100%**	0%	12%	**180 (18.0%)**	Principalement des activités ÉTUDES
C5	21%	2%	**100%**	33%	**126 (12.6%)**	Principalement des activités MAGASINAGE

Résultats du clustering: Test (2)



Contributions potentielles

Cette recherche vise à contribuer à –

- **l'enrichissement** du cadre de modélisation basée sur les activités par l'ajout d'une **procédure de segmentation;**
- Définir des **segments de population** qui ont des propriétés d'activités similaires, ce qui facilitera la **construction de distributions** (fréquence d'activités, heure de début, durée) adaptées à la prévision de la demande.

Merci!