Modèles à variables latentes, Fouille de données, EM et ses dérivés

Séminaire COSYS, Modélisation

Etienne Côme
COSYS/GRETTIA

16 décembre 2013

IFSTTAR

Plan

Contexte

Etienne Côme

- CR : GRETTIA, COSYS
- Analyse de données, Clustering, Apprentissage statistique,...
- Fouille de données de mobilité : ex VLS (Vélib',Velov,...)

Traces numériques de mobilité

De nombreux capteurs

Différentes formes : Trajectoires

Différentes formes : Points

Différentes formes : Origines-Destinations

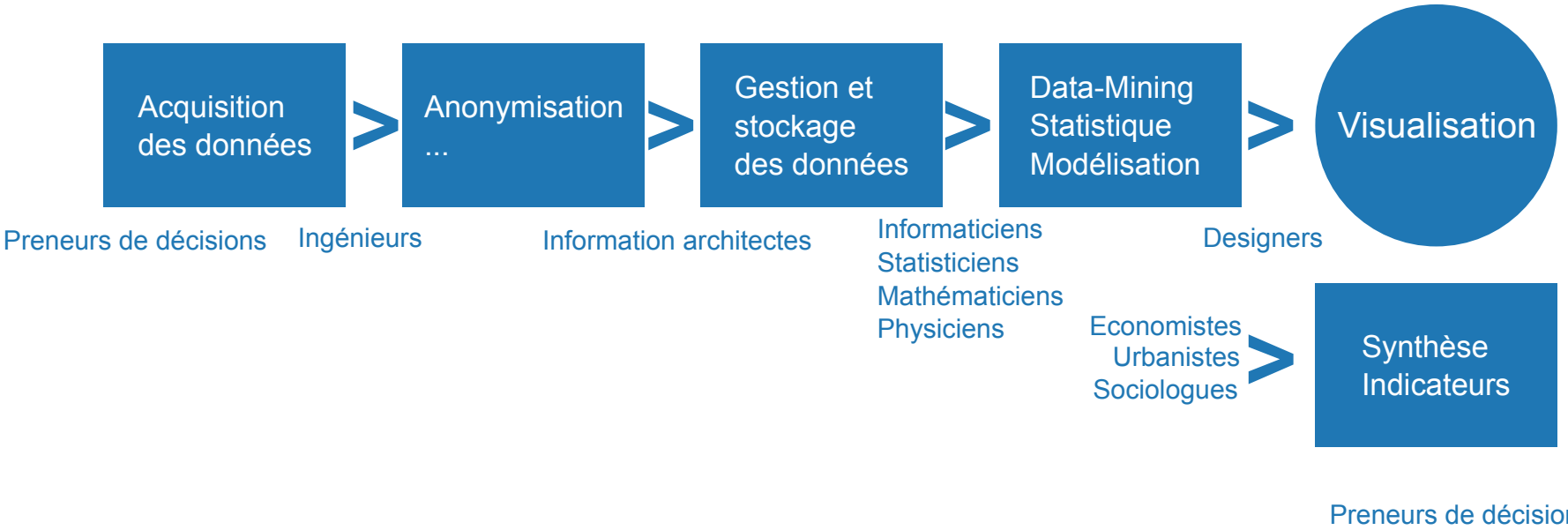

Nécessite de fouiller ! !
www.képtelenség.hu

Contexte

Problématique

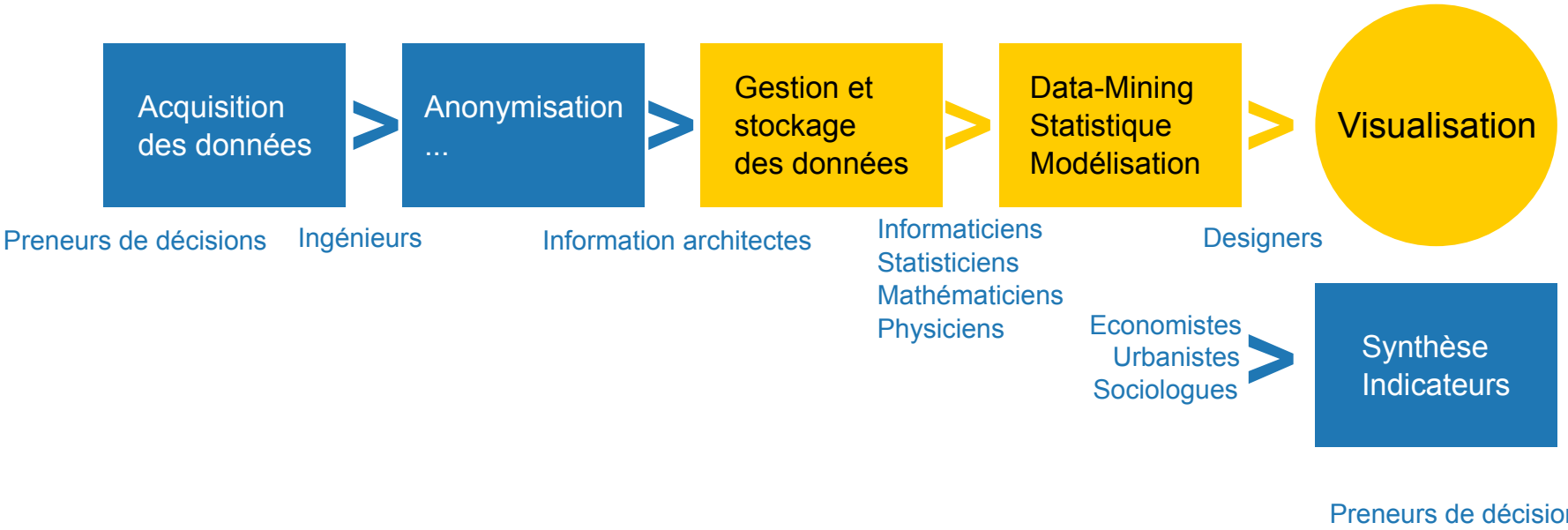
- Traces numériques de plus en plus nombreuses
- Générées en particulier lors de nos déplacements
  $\Rightarrow$ Pertinentes pour l'étude des mobilités
- Usage détourné + Volume important
  $\Rightarrow$ Outils de traitement et d'analyse
- Complémentarité / substitution enquêtes classiques

Métiers, compétences



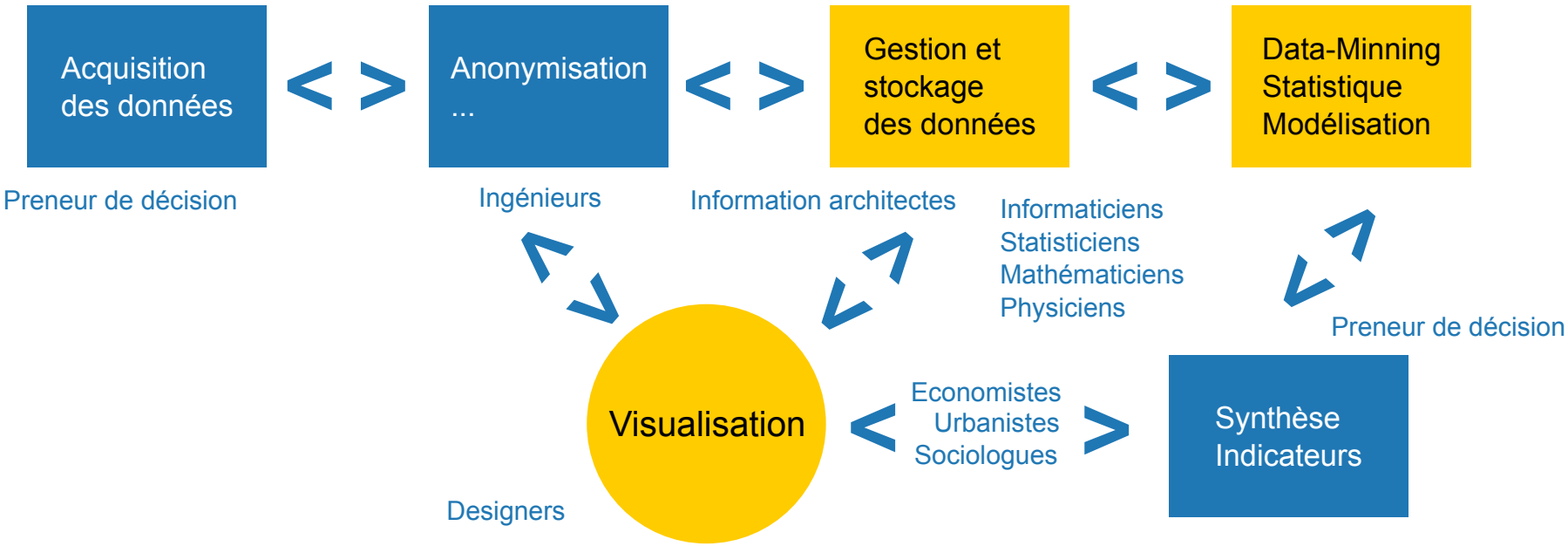
Mobilise différentes compétences, disciplines

Métiers, compétences



Nos focus, gestion stockage des données, fouille, visualisation

Métiers, compétences



Démarche non linéaire, importance du dialogue entre les acteurs.

Analyse exploratoire VLS : laisser parler les données

Méthodologie générale

- Utiliser des algorithmes de clustering (regroupement automatique) pour trouver des formes d'usage type du Vélib'
- Croiser les groupes ou clusters trouvés avec des données géographiques et socio-économiques de la ville
  $\Rightarrow$ Facteurs influents sur l'usage du système VLS.

1. Trouver des groupes de stations similaires
2. Trouver des groupes d'utilisateurs similaires
3. Segmenter / Résumer la dynamique temporelle globale du système

Apprentissage statistique : supervisé / non-supervisé

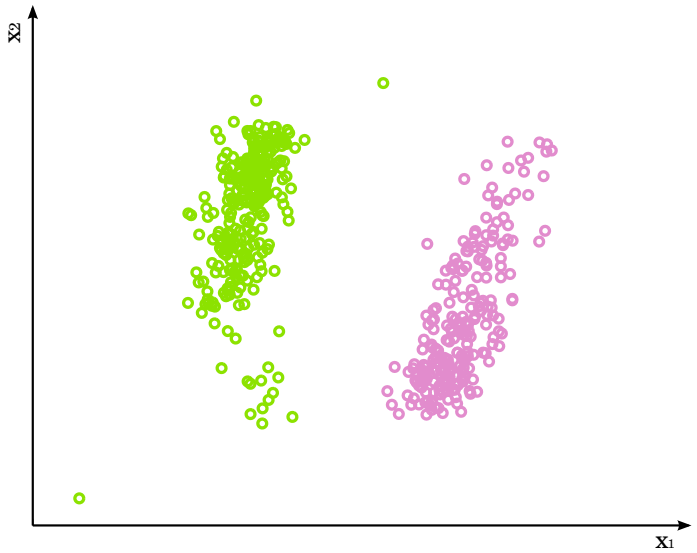


FIGURE 1: Création d'une base de données labellisées.

Apprentissage supervisé

Ensemble exemples de mesure $\mathbf{x}_i$
dont le mode de fonctionnement : y_i est connu de manière certaine.

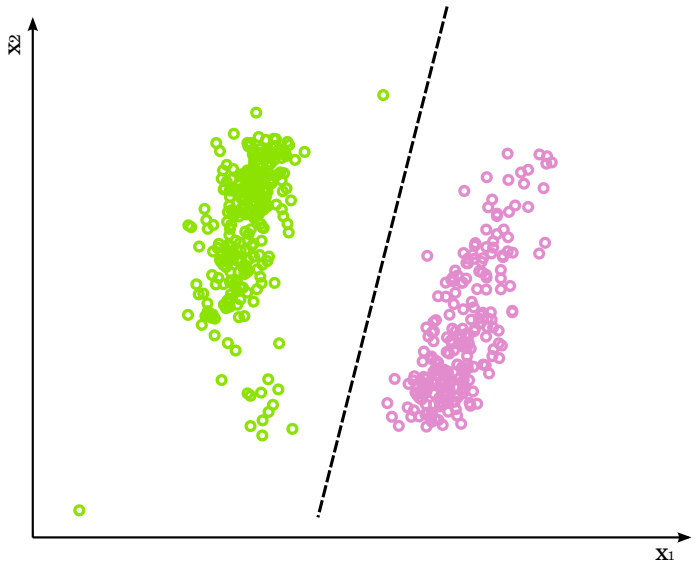


FIGURE 2: Construction de la frontière de décision.

Approche discriminative

Frontière permettant de séparer les individus des différentes classes ; estimer les paramètres de la loi conditionnelle $p(y|\mathbf{x})$

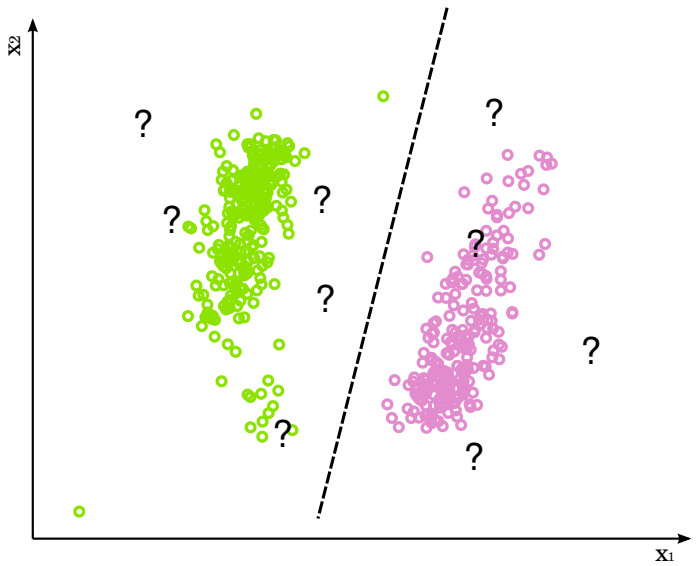


FIGURE 3: Classification de nouveaux points.

Classification de nouveaux points

Permet de déterminer le mode de fonctionnement le plus probable de toutes nouvelles mesures fournies en entrée $\arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x})$.

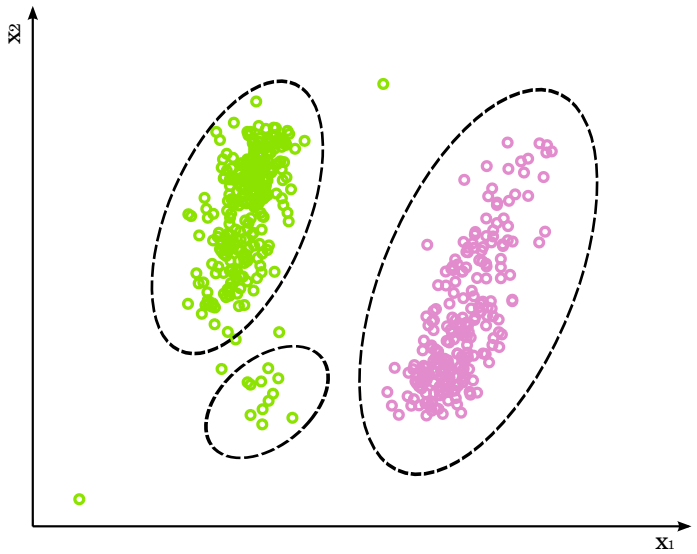


FIGURE 4: Modélisation de la loi jointe.

Approche générative

Modéliser et estimer la loi jointe $p(y, \mathbf{x})$;
permet ensuite de classer de nouvelles mesures $\arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x})$.

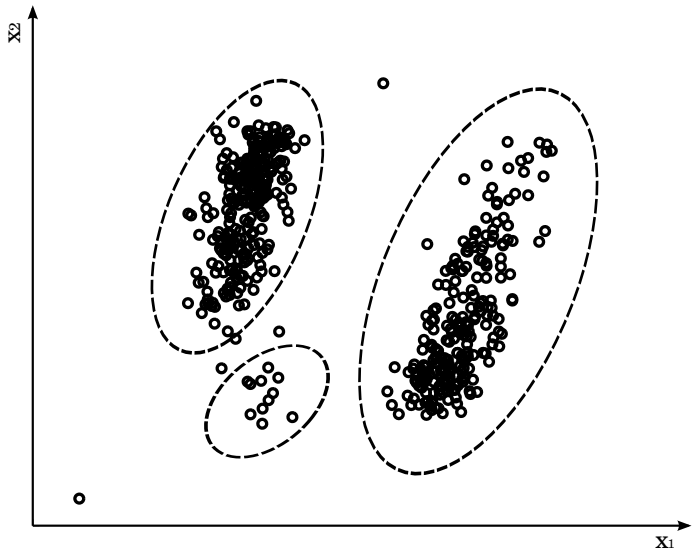


FIGURE 5: Modélisation de la loi jointe dans un contexte non supervisé (1).

Approche générative

Permet de travailler dans un contexte non supervisé ;
sans information sur la classe d'appartenance des exemples.

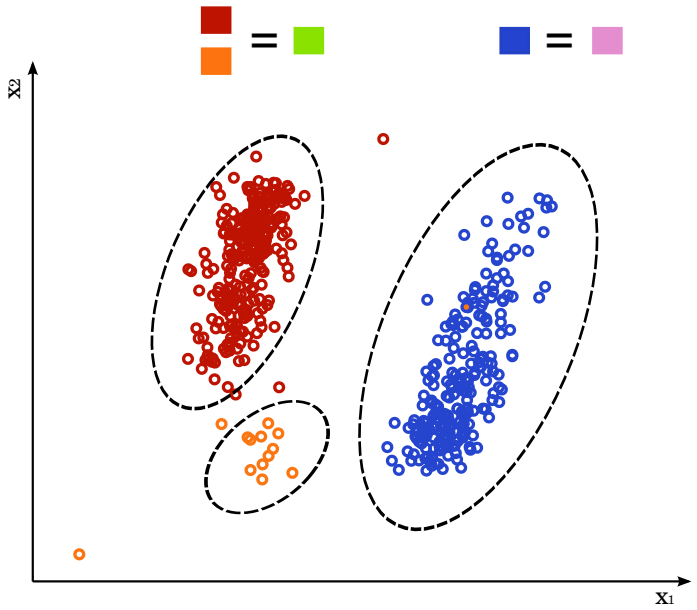


FIGURE 6: Modélisation de la loi jointe dans un contexte non supervisé (2).

Approche générative

Permet de travailler dans un contexte non supervisé ;
$\Rightarrow$ approche "model based clustering"

Modèles à variables latentes

Clustering // "Model based clustering"

Clustering

Recherche d'une structure de groupe
algorithme ad-hoc, cah, ...
$\Rightarrow$! importance de la distance utilisée

Model based Clustering : modèle de mélange

Recherche d'une structure de groupe
$\Rightarrow$! en utilisant un modèle statistique
$\Rightarrow$ faisant intervenir des variables latentes (inobservées)
$\Rightarrow$ approche générative :
définition d'une procédure de génération des données.

"*Tous les modèles sont faux mais certains sont utiles.*"

Georges Box.

Approche générative

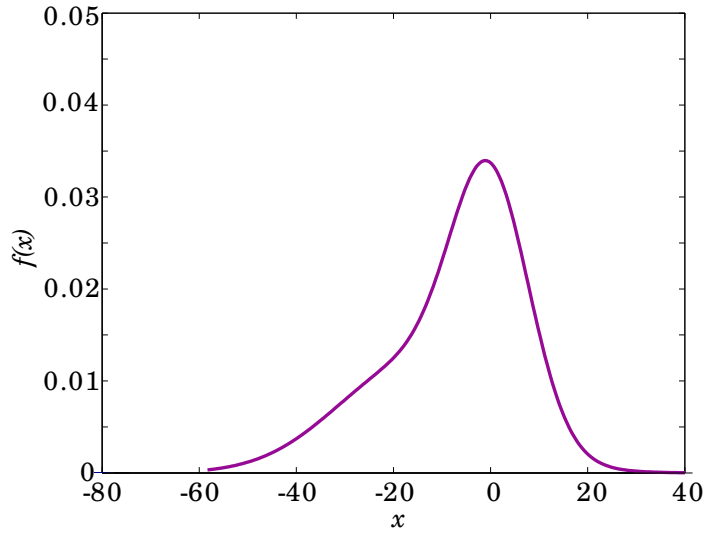


FIGURE 7: Exemple de densité pour un mélange de Gaussiennes.

Modèle de mélange, simulation :

1. Tirage aléatoire de la classe
2. Suivant la classe tirée ; tirage aléatoire d'une mesure

Approche générative

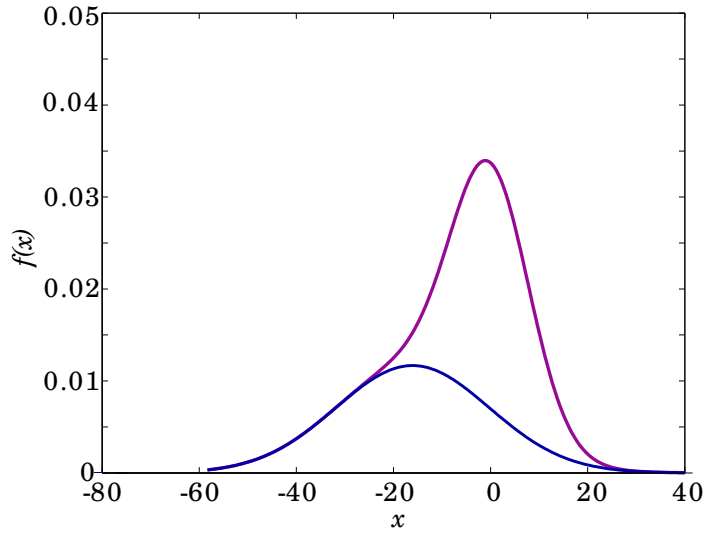


FIGURE 8: Exemple de densité pour un mélange de Gaussiennes.

Modèle de mélange, simulation :

1. Tirage aléatoire de la classe
2. Suivant la classe tirée ; tirage aléatoire d'une mesure

Approche générative

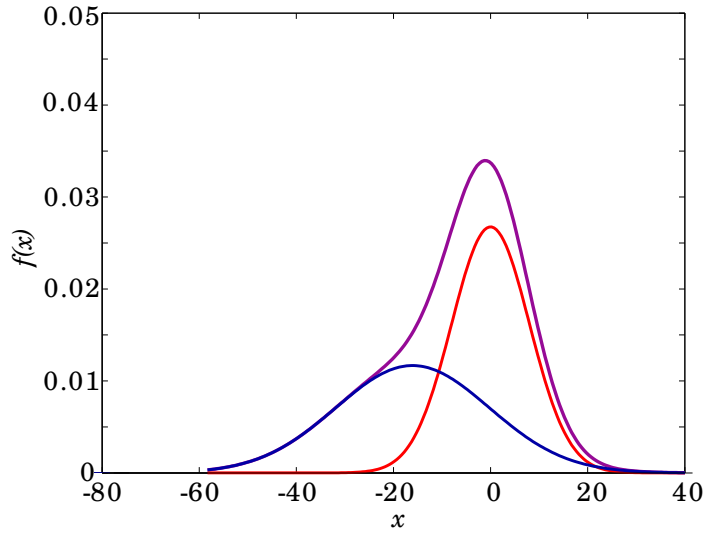
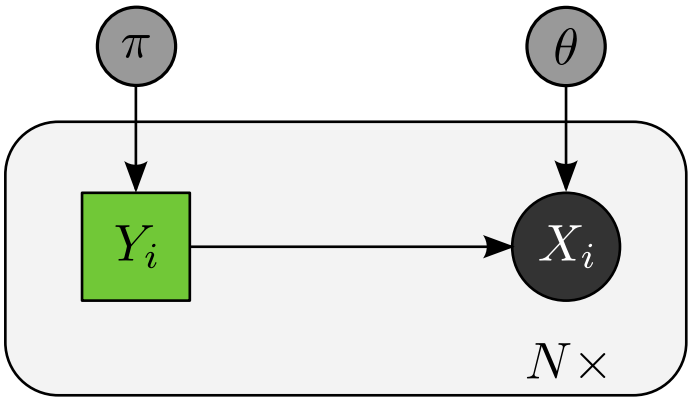


FIGURE 9: Exemple de densité pour un mélange de Gaussiennes.

Modèle de mélange, simulation :

1. Tirage aléatoire de la classe
2. Suivant la classe tirée ; tirage aléatoire d'une mesure

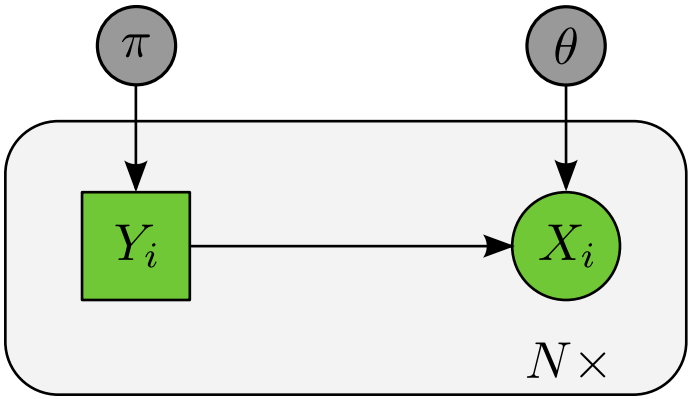
Schéma de génération des données



Tirage de la classe :

$$Y_i \sim \mathcal{M}(1, \boldsymbol{\pi}) \tag{1}$$

Schéma de génération des données

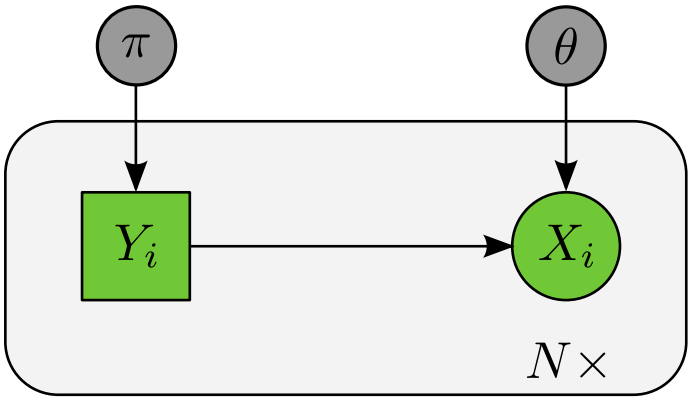


Connaissant la classe, tirage de l'observation :

$$\mathbf{X}_i | Y_{ik} = 1 \sim \mathcal{F}(\theta) \tag{2}$$

Suivant hypothèses sur $\mathcal{F} \rightarrow$ diversité de modèles

Schéma de génération des données



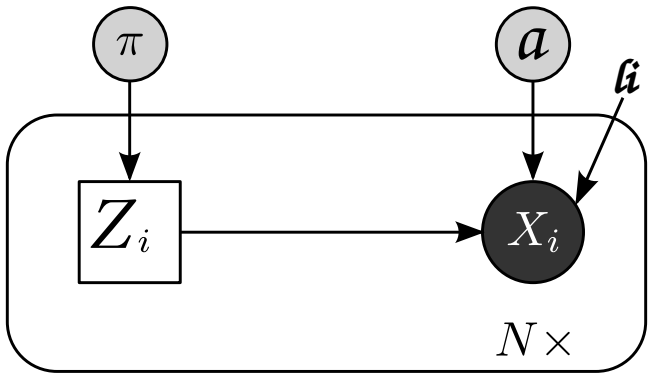
Connaissant la classe, tirage de l'observation :

$$\mathbf{X}_i | Y_{ik} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

$\mathcal{N}$ mélange de Gaussiennes
+ hypothèses supplémentaires sur forme de $\boldsymbol{\Sigma}$

Analyse de textes : Mixture of unigrams

Mixture of unigrams



Modèle de mélange de multinomiales

$$\begin{aligned} Z_i &\sim \mathcal{M}(1, \pi) \\ \mathbf{X}_i | Z_{ik} = 1 &\sim \mathcal{M}(l_i, \alpha_k) \end{aligned} \tag{4}$$

$\mathcal{M}$ Multinomiale, "mixture of unigrames" pour l'analyse de textes, l_i longueur du texte.

Exemple de données simulées

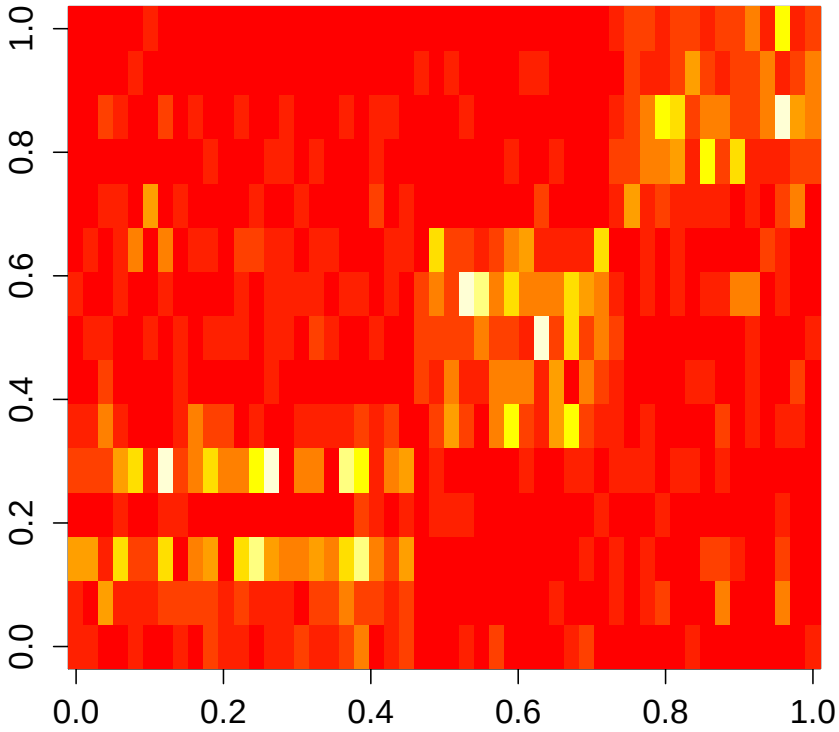


FIGURE 10: 50 documents, 15 mots, structure forte.

Analyse de textes : Latent Dirichlet Allocation

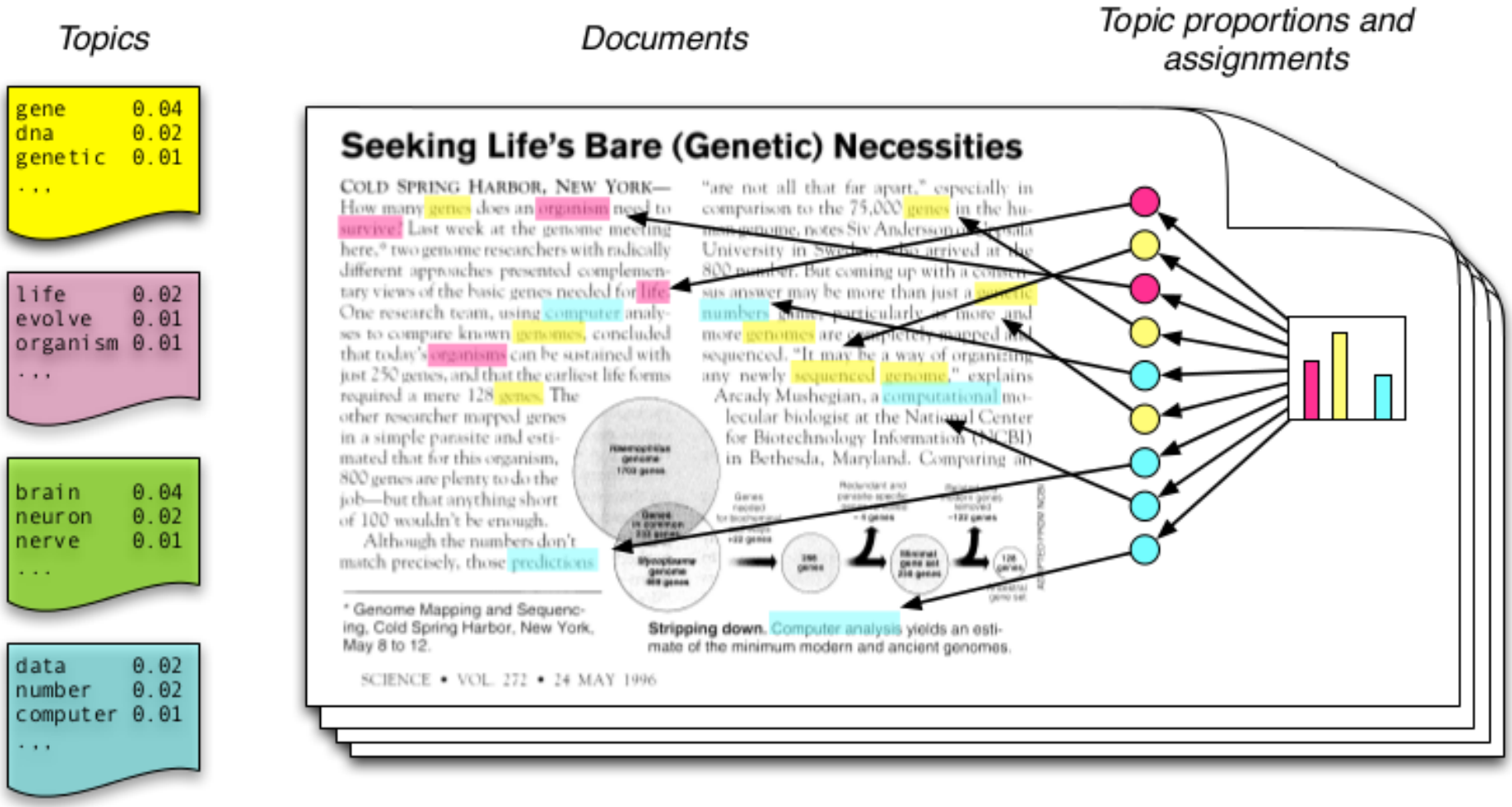


FIGURE 11: Illustration du schéma génératif de LDA, source *D. Blei*

LDA

- pour chaque topics $a \in \{1, \ldots, N_a\}$ tirer une distribution sur le vocabulaire (Dirichlet) :

$$\Lambda_a \sim \mathcal{D}(\beta) \tag{5}$$

- pour chaque documents $t \in \{1, \ldots, N_t\}$:
  1. tirer ses proportions de topics :

$$\pi_t \sim \mathcal{D}(\alpha) \tag{6}$$

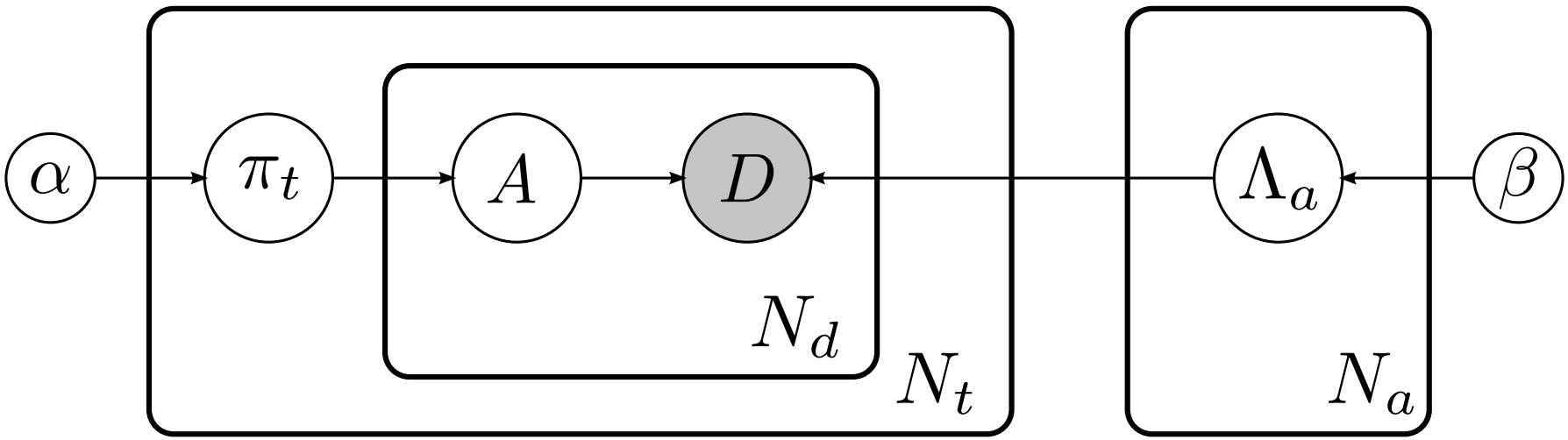
  2. pour chaque mots de t :
     - tirer le topic A du mots :

$$A \sim \mathcal{M}(1, \pi_t) \tag{7}$$

     - tirer un mots du vocabulaire D (couple O/D) conditionellement au topic :

$$D \sim \mathcal{M}(1, \Lambda_A) \tag{8}$$

LDA



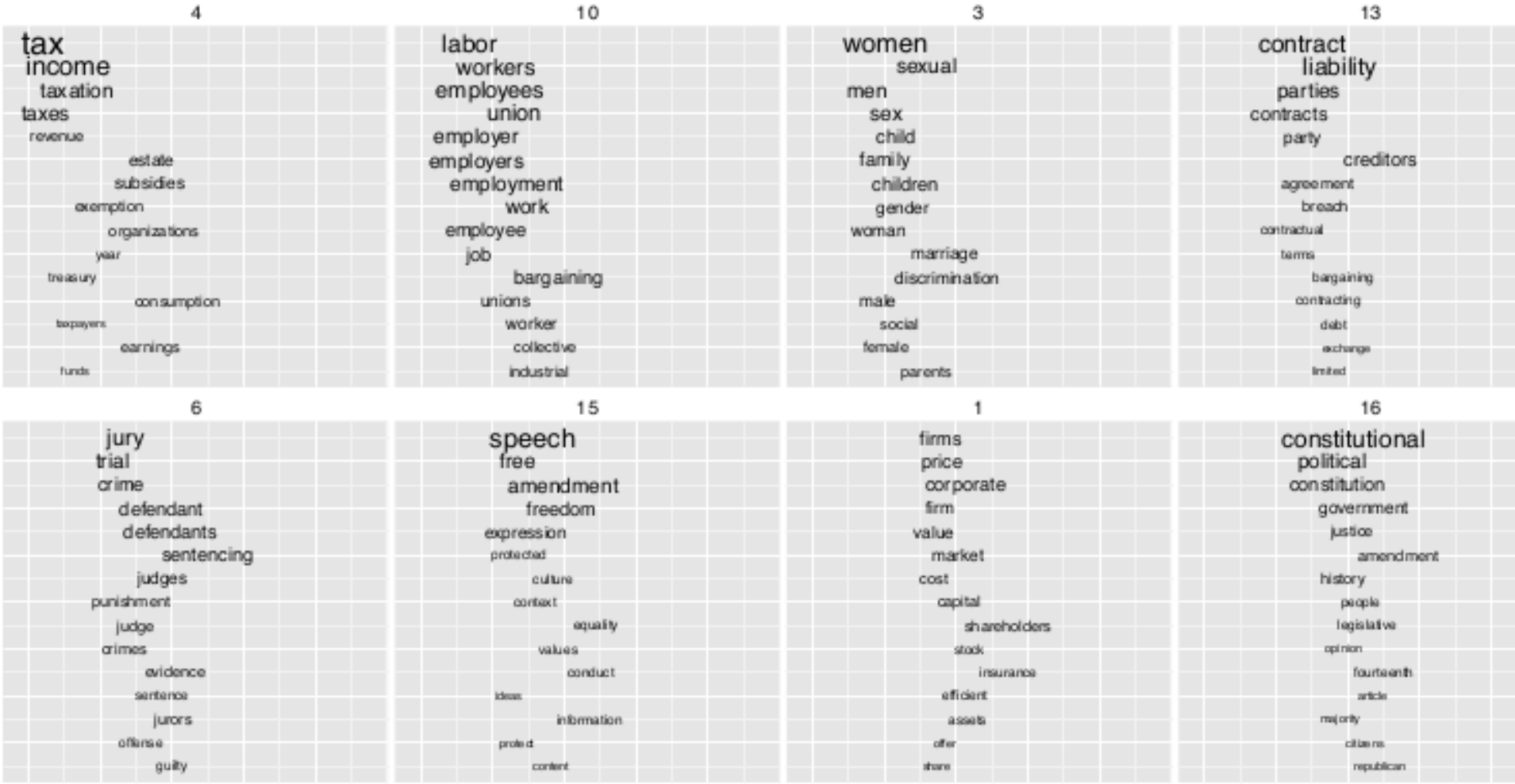
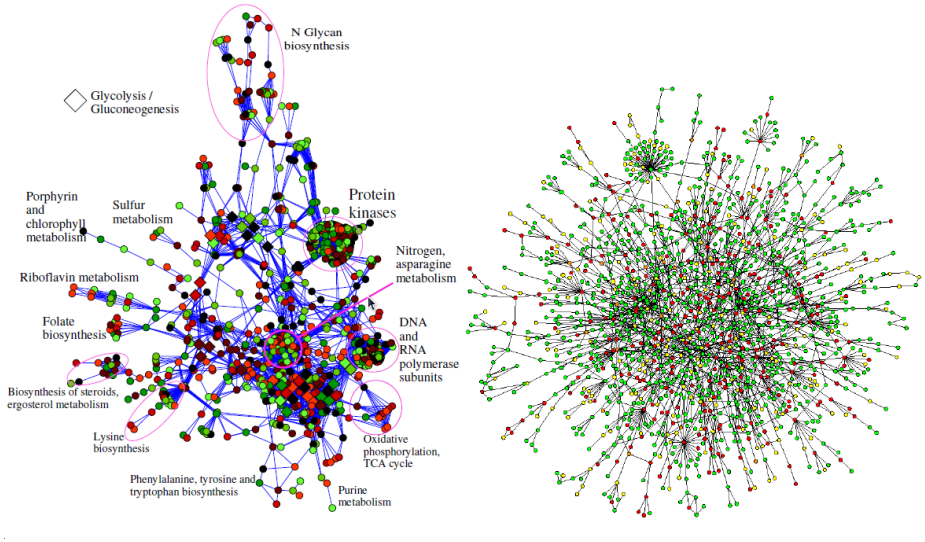


FIGURE 12: Exemple de topics, *"Yale law journal"*, source *D. Blei*

Analyse de graphe : Modèle de mélange d'Erdos Renyi

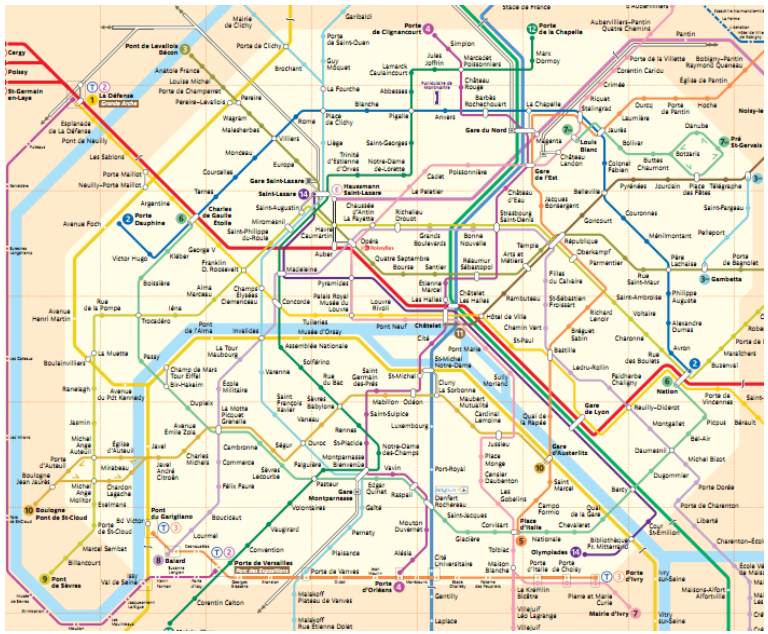
Beaucoup de domaines d'application

- réseaux routiers, biologiques, sociaux, ...
- analyse de données dans $\mathbb{R}^p$ en utilisant un noyau Gaussien ou $k - ppv$
- ...



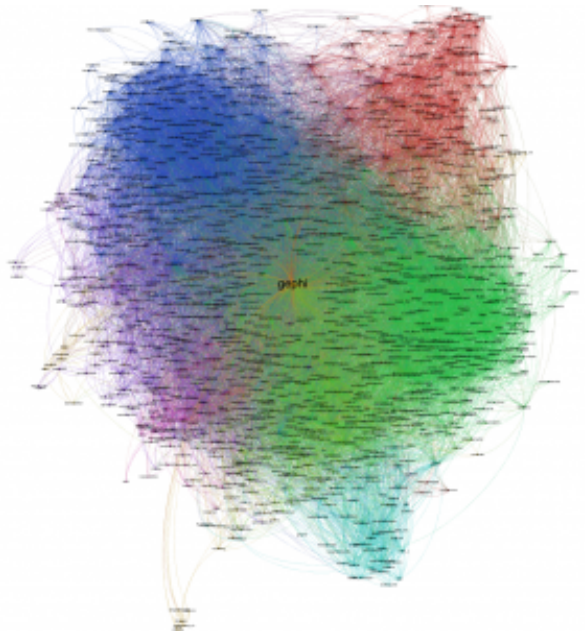
Beaucoup de domaines d'application

- réseaux routiers, biologiques, sociaux, ...
- analyse de données dans $\mathbb{R}^p$ en utilisant un noyau Gaussien ou $k - ppv$
- ...



Beaucoup de domaines d'application

- réseaux routiers, biologiques, sociaux, ...
- analyse de données dans $\mathbb{R}^p$ en utilisant un noyau Gaussien ou $k - ppv$
- ...



Graphes, représentations

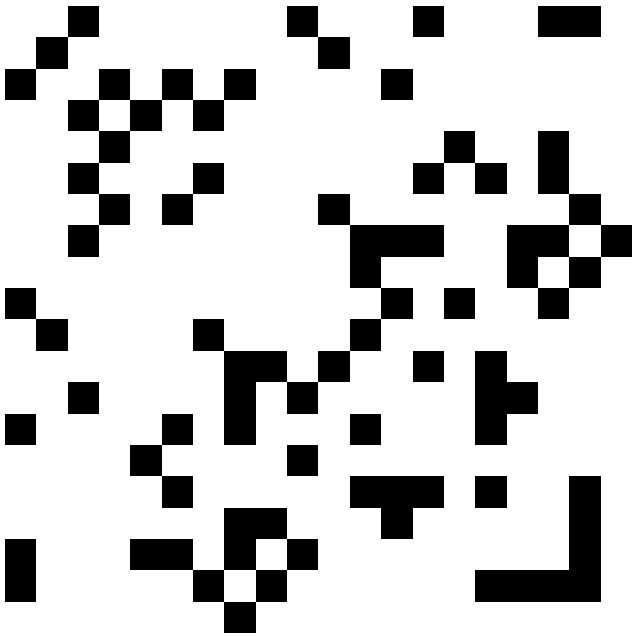
- Matrice d'adjacence A :

$$A : \begin{cases} A_{ij} = 1, \text{ si } i \sim j \\ A_{ij} = 0, \text{ sinon.} \end{cases}$$

Graphes, représentations

- Matrice d'adjacence A :

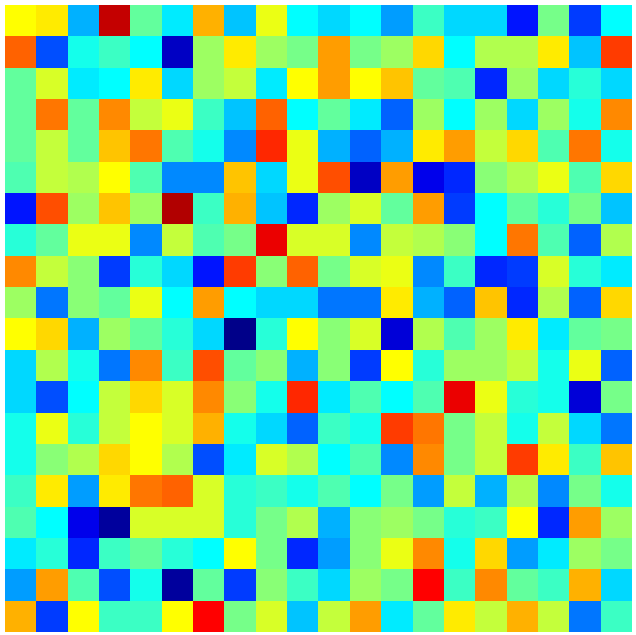
$$A : \begin{cases} A_{ij} = 1, \text{ si } i \sim j \\ A_{ij} = 0, \text{ sinon.} \end{cases}$$



Graphes, représentations

- Matrice d'adjacence A :

$$A : \begin{cases} A_{ij} = 1, \text{ si } i \sim j \\ A_{ij} = 0, \text{ sinon.} \end{cases}$$



Modèle de mélange d'Erdos Renyi

Variables :

- $X_{ij} \in \{0, 1\}$ variable binaire encodant la présence ou l'absence d'un liens entre i et j :

$$x_{ij} = \begin{cases} 1, \text{ si il existe un liens entre } i \text{ et } j \\ 0, \text{ sinon.} \end{cases} \tag{9}$$

- $Z_j \in \{0, 1\}^K$ sont des variables latentes, décrivant l'appartenance de j à un des K clusters possibles :

$$z_{jk} = 1, \text{ si } j \text{ appartient au cluster } k. \tag{10}$$

Modèle de mélange d'Erdos Renyi

Modèle génératif :

1. tirer le groupe de chaque noeud suivant les proportions γ
2. ajouter un lien entre i et j avec une probabilité π_{kl} si i appartient au cluster k et j appartient au cluster l.

$$Z_j \overset{i.i.d}{\sim} \mathcal{M}(1,\gamma), \quad \forall j \in \{1,\dots,N\} \tag{11}$$

$$X_{ij}|Z_{ik}Z_{jl}=1 \overset{i.i.d}{\sim} \mathcal{B}(\pi_{kl}), \quad \forall i,j \in \{1,\dots,N\}, \tag{12}$$

Paramètres :

- γ : proportions, exemple $\gamma = (0.1, 0.2, 0.6, 0.1)$
- π : matrice de liens, exemple :

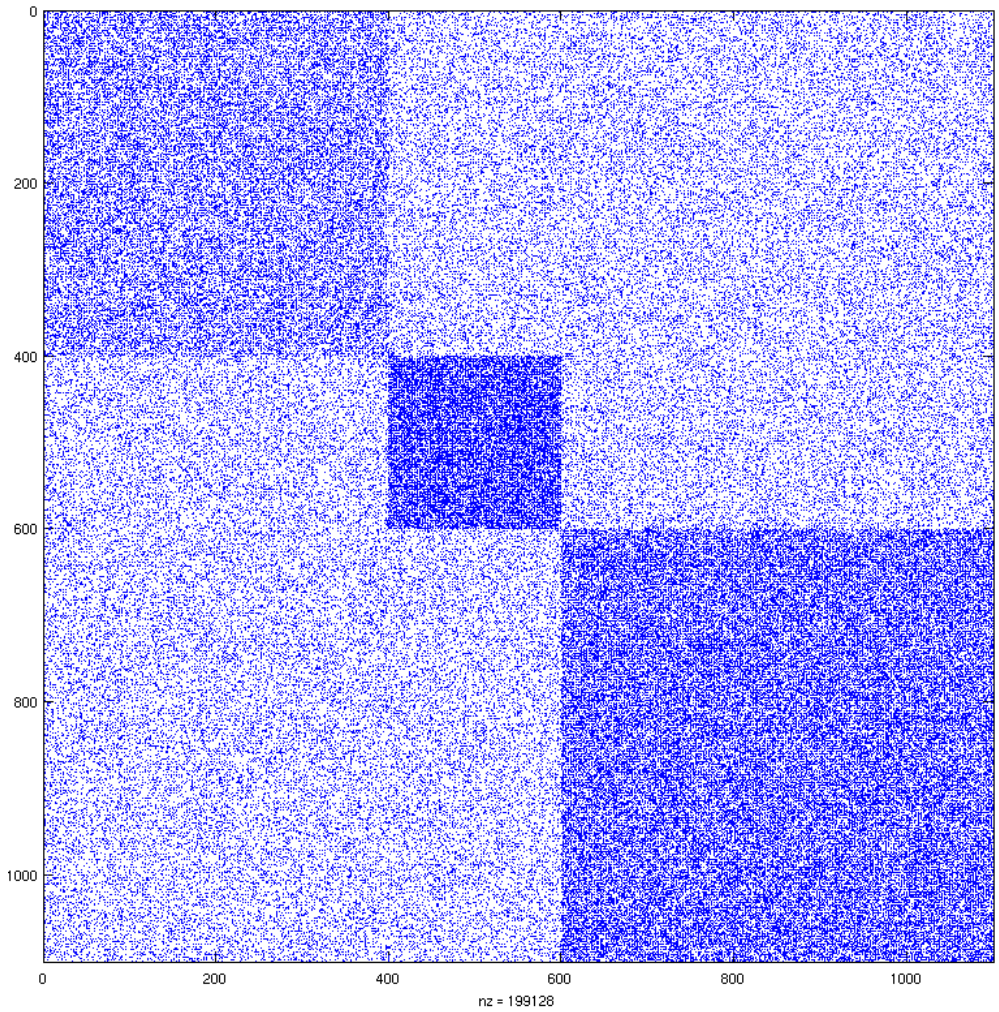
$$\pi = \begin{pmatrix} 0.1 & 0.01 & 0.01 & 0.005 \\ 0.005 & 0.2 & 0.01 & 0.01 \\ 0.005 & 0.001 & 0.1 & 0.01 \\ 0.005 & 0.001 & 0.01 & 0.3 \end{pmatrix}.$$

Recherche de communauté :

$$\pi = \begin{pmatrix} \alpha_1 & \epsilon & \epsilon & \epsilon \\ \epsilon & \alpha_2 & \epsilon & \epsilon \\ \epsilon & \epsilon & \alpha_3 & \epsilon \\ \epsilon & \epsilon & \epsilon & \alpha_4 \end{pmatrix},$$

avec $\alpha >> \epsilon$.

Exemple de réalistation



Bi-clustering, graphes bi-partites

Modèle génératif :

- structures de groupes différentes en lignes / colonnes
- nbr de lignes $\neq$ nbr de colonnes

$$
\begin{aligned}
Z_i^c &\overset{i.i.d}{\sim} \mathcal{M}(1, \gamma^c), \quad \forall i \in \{1, \dots, N_c\} \\
Z_j^r &\overset{i.i.d}{\sim} \mathcal{M}(1, \gamma^r), \quad \forall j \in \{1, \dots, N_r\} \\
X_{ij} | Z_{ik}^c Z_{jl}^r = 1 &\overset{i.i.d}{\sim} \mathcal{B}(\pi_{kl}), \quad \forall i, j
\end{aligned} \tag{13}
$$

Tri-clustering

Modèle génératif :

- extension aux tenseurs :

$$
\begin{aligned}
Z_i^c &\overset{i.i.d}{\sim} \mathcal{M}(1, \gamma^c), \quad \forall i \in \{1, \dots, N_c\} \\
Z_j^r &\overset{i.i.d}{\sim} \mathcal{M}(1, \gamma^r), \quad \forall j \in \{1, \dots, N_r\} \\
Z_h^t &\overset{i.i.d}{\sim} \mathcal{M}(1, \gamma^t), \quad \forall h \in \{1, \dots, N_t\} \\
X_{ijh} | Z_{ik}^c Z_{jl}^r Z_{hm}^r = 1 &\overset{i.i.d}{\sim} \mathcal{B}(\pi_{klm}), \quad \forall i, j, h
\end{aligned} \tag{14}
$$

Estimation : Algorithme EM

Algorithme EM

Les modèles de mélange

paramètres $\psi = (\pi, \theta)$,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \theta_k) \quad (15)$$

$$\mathcal{L}(\psi; \mathbf{X}^{ns}) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i, \theta_k)\right) \quad (16)$$

$f_k(., \theta_k)$ densité paramétrique sur $\mathcal{X}$.

Maximisation / paramètres (Maximum de vraisemblance)

Problème d'optimisation non convexe
Solution élégante : l'algorithme EM

Algorithme EM

Décomposition de la log vraisemblance :

Introduction de la variable latente z : $p(x) = \frac{p(x,z)}{p(z|x)}$
Permet de réécrire la fonction de log vraisemblance :

$$\mathcal{L}(\psi; \mathbf{X}^{ns}) = \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(\pi_k f(\mathbf{x}_i; \theta_k)\right)}_{Q(\psi,\psi^{(q)})} - \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(t_{ik}\right)}_{H(\psi,\psi^{(q)})}, \quad (17)$$

—

avec :

$$t_{ik}^{(q)} = \mathbb{E}_{\psi^{(q)}}[z_{ik}|\mathbf{x}_i] = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \theta_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \theta_{k'}^{(q)})}. \quad (18)$$

Algorithme EM

—

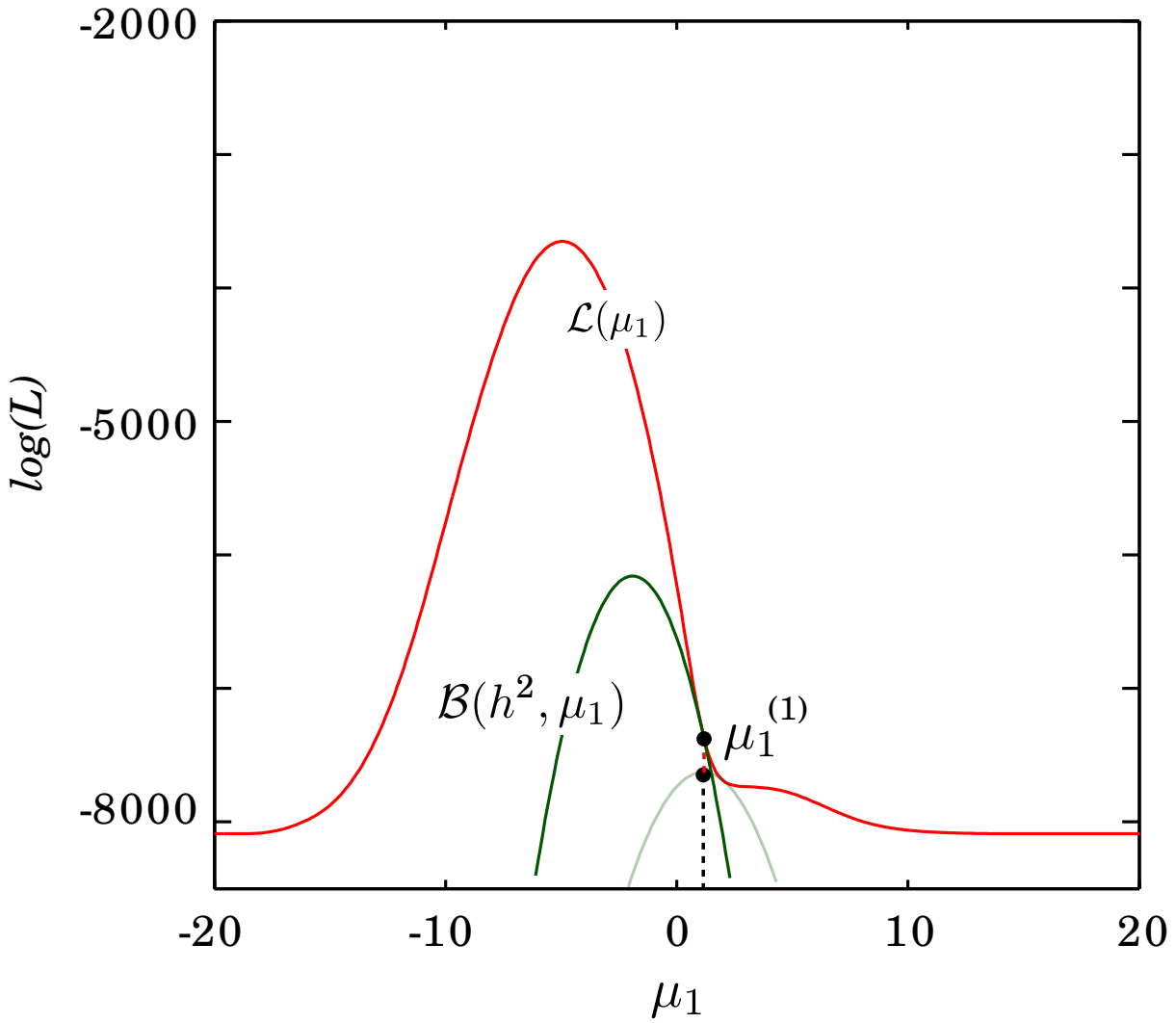
Initialisation : $\psi^{(0)}$

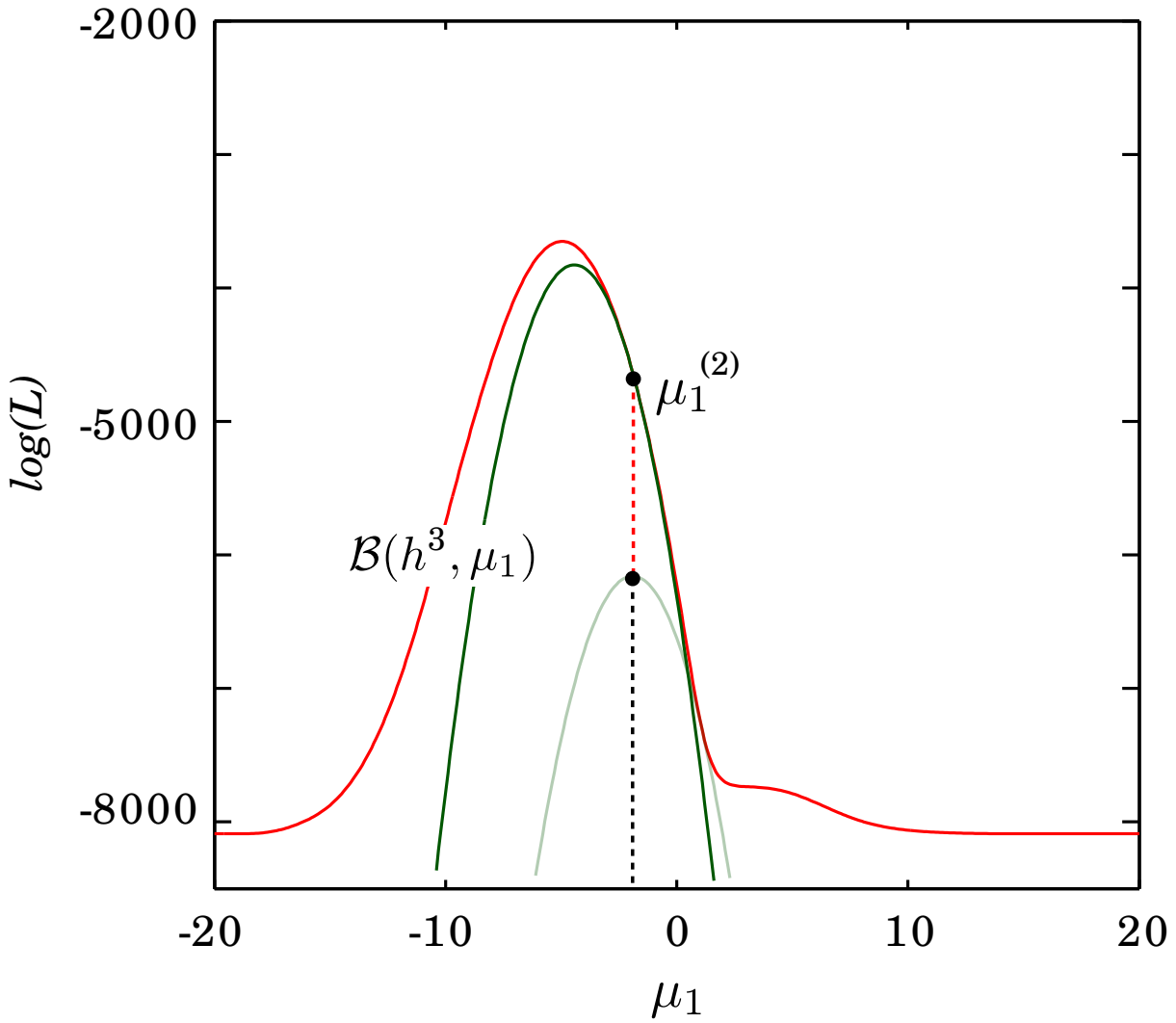
Cercle vertueux :

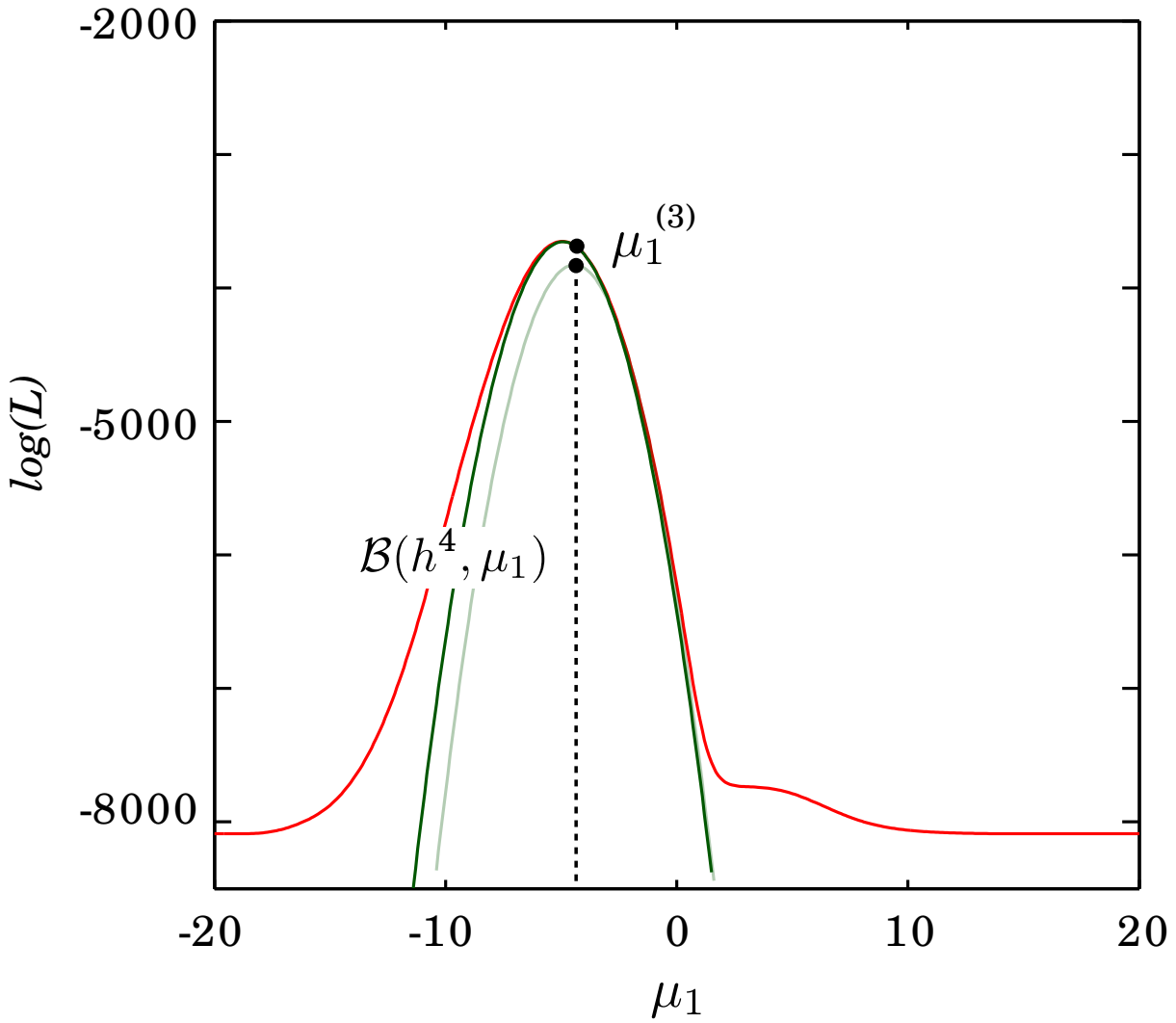
- **Espérance :** Calcul des $t_{ik}^{(q)}$
- **Maximisation :** $\psi^{(q+1)} = \arg\max_{\boldsymbol{\psi}} Q(\psi, \psi^{(q)})$
  Solution analytique pour les proportions : $\pi_k^{(q+1)} = \frac{1}{N} \sum_{i=1}^{N} t_{ik}^{(q)}$.
  Solution analytique pour les θ_k si lois classiques
  car $\sum$ en dehors du log.

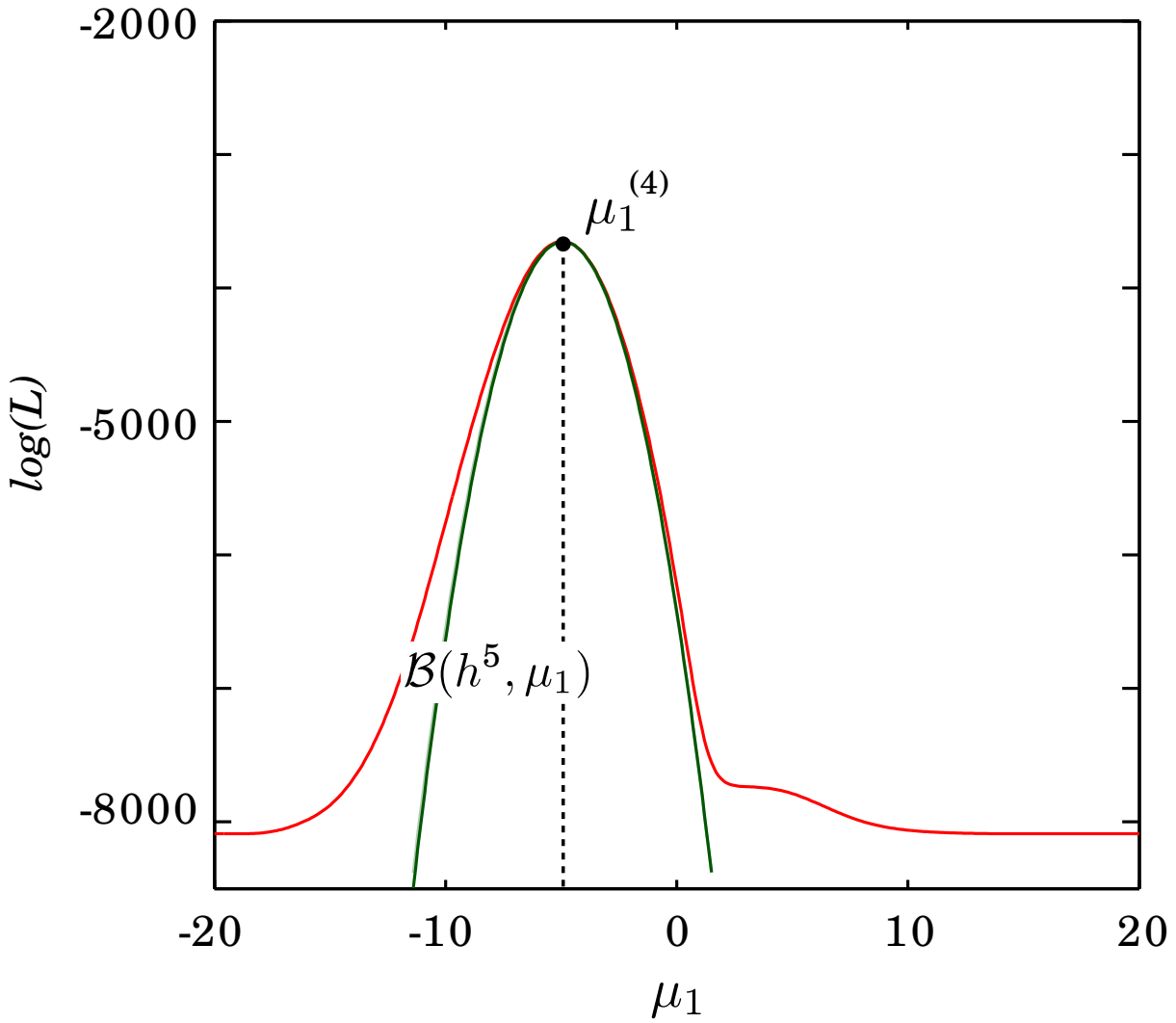
Convergence

$\Rightarrow$ maximisation de $Q(\psi, \psi^{(q)})$ suffisante pour faire croître la vraisemblance.


-2000
-5000
-8000
log(L)
-20
-10
0
10
20
μ_1
$\mathcal{L}(\mu_1)$
$\mathcal{B}(h^2, \mu_1)$
$\mu_1^{(1)}$


-2000
-5000
-8000
log(L)
-20
-10
0
10
20
μ_1
$\mu_1^{(2)}$
$\mathcal{B}(h^3, \mu_1)$


-2000
-5000
-8000
log(L)
-20
-10
0
10
20
μ_1
$\mu_1^{(3)}$
$\mathcal{B}(h^4, \mu_1)$


-2000
-5000
-8000
log(L)
-20
-10
0
10
20
μ_1
$\mu_1^{(4)}$
$\mathcal{B}(h^5, \mu_1)$

Estimation : au delà de EM

Variations, Extensions

Variations

- seuillage des probabilités a posteriori (après étape E) : CEM, k-means
- Etape E stochastique : SEM,MCEM,...
- Etape M partielle : MEM
- Online EM
- Recuit simulé : SAEM, DAEM
- ...

Variations, Extensions

Variationel : VEM, VBEM

Problème :

- variables latentes pas indépedantes | aux observations :
  $\Rightarrow$ Champs de markov cachés,
  $\Rightarrow$ Modéle de mélange d'Erdos-Renyi, ...
- Approximer la distribution a posteriori par une autre $R(\mathbf{Z})$
- Utiliser la décomposition :

$$L(\theta, \mathbf{X}) = \mathcal{L}\left(R(\mathbf{Z}), \theta\right) + \mathcal{KL}\left(R(\mathbf{Z}) || \mathbb{P}(\mathbf{Z}|\mathbf{X})\right), \tag{19}$$

avec $\mathcal{L}(R(\mathbf{Z}), \theta)$ une borne inférieure *ELBO* :

$$\mathcal{L}(R(\mathbf{Z}), \theta) = \mathcal{H}\left(R(\mathbf{Z})\right) + \sum_{\mathbf{Z}} R(\mathbf{Z}) \log\left(\mathbb{P}(\mathbf{X}, \mathbf{Z})\right) \tag{20}$$

Variations, Extensions

Variationel : VEM, VBEM

Problème :

- variables latentes pas indépedantes | aux observations :
  $\Rightarrow$ Champs de markov caché,
  $\Rightarrow$ Modéle de mélange d'Erdos-Renyi, ...
- $R(\mathbf{Z})$ choisie pour pouvoir calculer $\mathcal{L}(R(\mathbf{Z}), \theta)$,
- Approximation en champs moyen :

$$R(\mathbf{Z}) = \prod_{i,k} \tau_{ik}^{Z_{ik}} \approx \text{indépendantes}$$

Optimisation alternée de $\mathcal{L}(R(\mathbf{Z}), \theta)$ // τ_{ik} et θ. point fixe sur les τ_{ik}.

EM

Autres points

- Extensions Bayesiennes
- MCMC
- Sélection de modèle
- Sélection de K
- ...

Exemple 1 :

Clustering de stations // profils d'usage temporels

Clustering de station // profils d'usage temporels

Objectifs :

- Stations décrites par les dynamiques de flux entrants et sortants
  - X^{out}_{sdt} : # de vélos pris à la station s le jour d à l'heure t
  - X^{in}_{sdt} : # de vélos déposés à la station s le jour d à l'heure t

$$\mathbf{X}_{sd} = (X^{in}_{sd1}, \ldots, X^{in}_{sd24}, X^{out}_{sd1}, \ldots, X^{out}_{sd24})$$

- Prise en compte des jours de semaine / week end.
- Croiser les résultats avec d'autres variables explicatives : population, emplois, loisirs, ...
- Analyse réalisée avec 8 groupes
  (bon compromis : interprétations/attache aux données)

Modèle génératif

- X_{sdt} (observée) : Nbr de vélos arrivant/partant
- Z_s (latente) : cluster de la station s
- W_d (observée) : cluster des jours (week end / semaine)

Mélange de poissons

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

$$X_{sdt} | \{Z_{sk} = 1, W_{dl} = 1\} \sim \mathcal{P}(\alpha_s \lambda_{klt})$$

+ contraintes $\sum_{l,t} D_l \lambda_{klt} = DT, \forall k \in \{1, \ldots, K\}$, avec D_l nbr de jours du cluster l.

Related work

[Rau et al., 2011], [Witten et al., 2010], [Govaert and Nadif, 2010]...

Vraisemblance

$$g(X_s) = \sum_{k=1}^{K} \pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}} \tag{21}$$

où $po(., \lambda)$: densité d'une loi de Poisson de paramètre λ

La log-vraisemblance à maximiser :

$$L(\mathbf{\Theta}; \mathbf{X}, \alpha, \mathbf{W}) = \sum_{s} \log \left(\sum_{k} \pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}} \right) \tag{22}$$

Les paramètres du modèle à estimer sont $\boldsymbol{\pi}$ et $\boldsymbol{\lambda}$.

L'algorithme EM

Vraisemblance complétée

$$Lc(\mathbf{\Theta}; \mathbf{X}, \mathbf{Z}, \alpha, \mathbf{W}) = \sum_{s,k} Z_{sk} \log \left(\pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}} \right) \quad (23)$$

où Z est un vecteur inconnu.

L'algorithme EM

Etape E

Dans l'étape E, on s'intéresse à l'espérance de la vraisemblance complétée, donnée par :

$$Ec(\Theta; \alpha, \lambda) = \sum_{s,k} t_{sk} \log \left(\pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}} \right) \tag{24}$$

où les t_{sk} sont définis par :

$$t_{sk} = \frac{\pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}}}{\sum_k \pi_k \prod_{d,t,l} po(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}}} \tag{25}$$

L'algorithme EM

Etape M

- α_s : Activité moyenne de la station

$$\hat{\alpha}_s = \frac{1}{DT} \sum_{d,t} X_{sdt}, \tag{26}$$

- λ_{klt} : Activité de la tranche horaire t pour le cluster k, en week end ou en semaine (cluster temporel l)

$$\hat{\lambda}_{klt} = \frac{1}{\sum_s t_{sk} \alpha_s \sum_d W_{dl}} \sum_{s,d} t_{sk} W_{dl} X_{sdt} \tag{27}$$

- π_k : porportion de la classe k, $\hat{\pi}_k = \frac{1}{N} \sum_s t_{sk}$

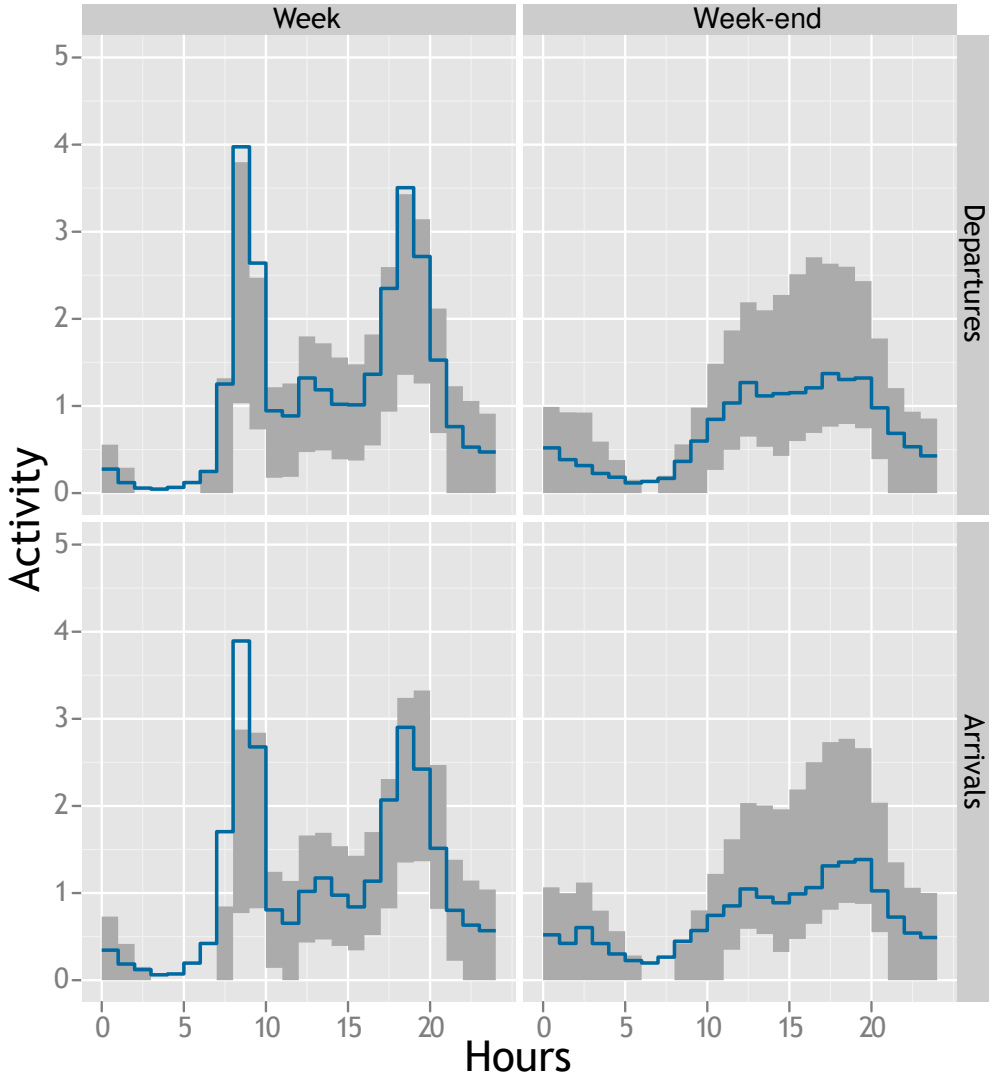
Analyse des résultats

Sorties

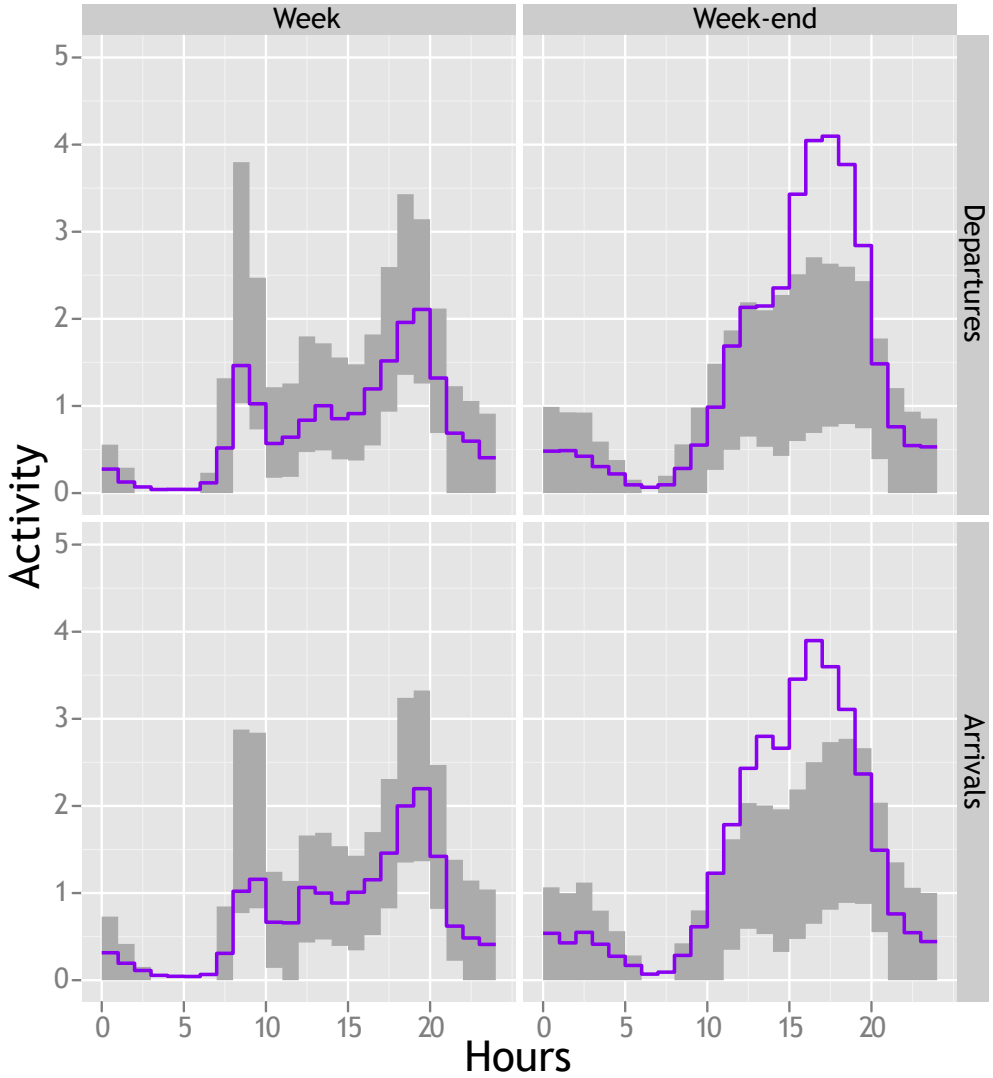
- Z_s : cluster des stations
- λ_k : profils temporels de la classe k
- α_s : effets stations

Pôles multimodaux

Pôles multimodaux

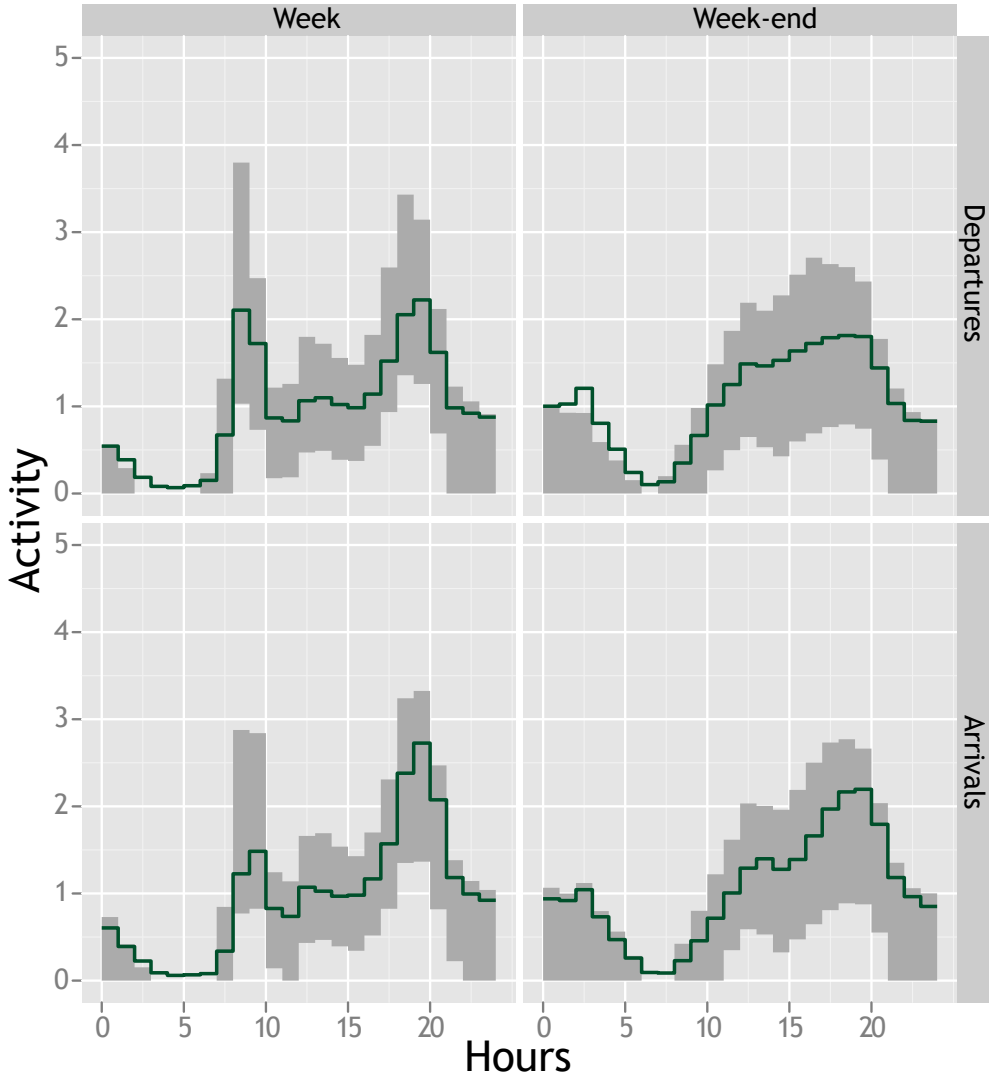


Parcs

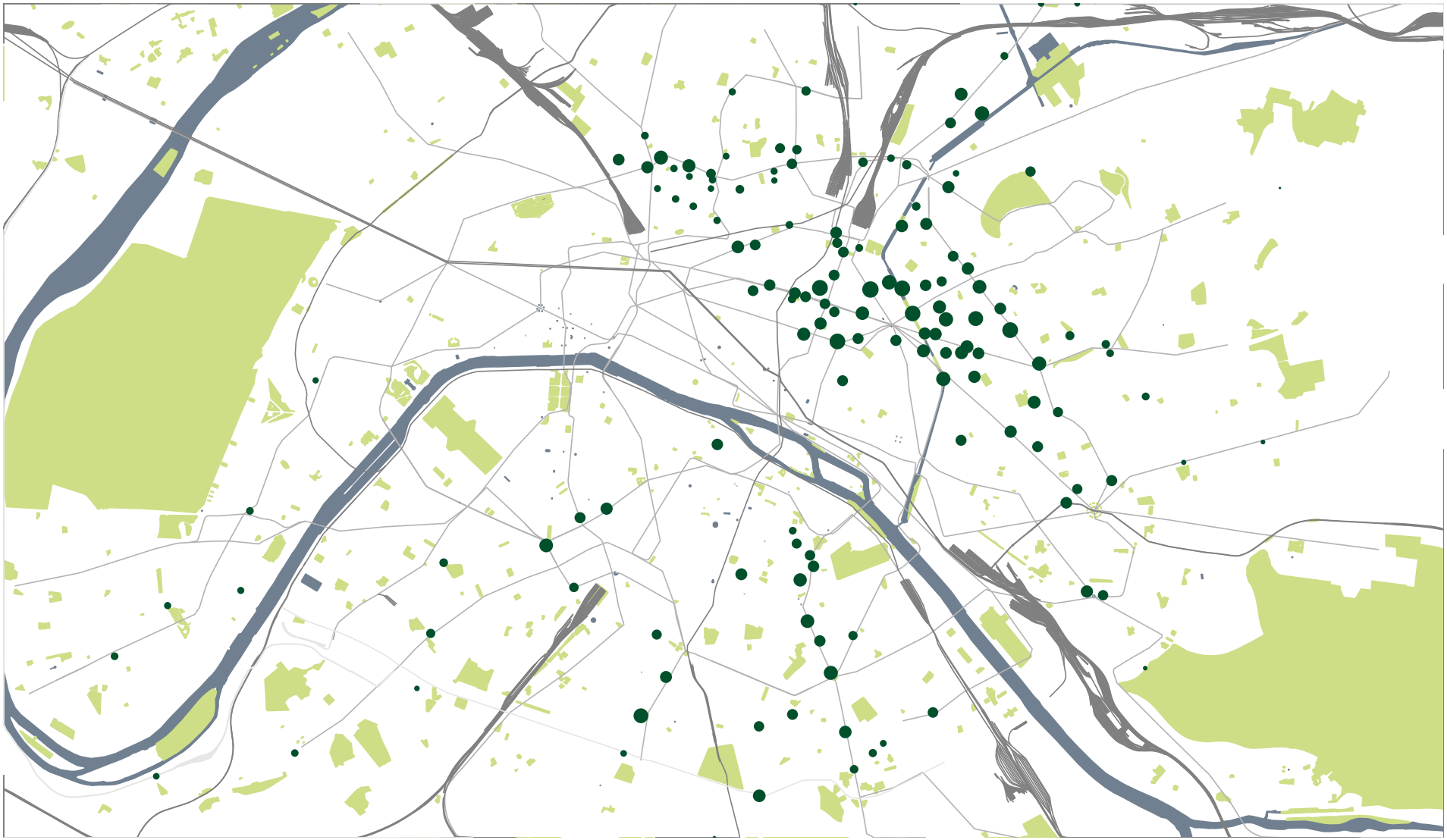

Week
Week-end
Departures
Arrivals
Activity
Hours
0
1
2
3
4
5
0
5
10
15
20

Parcs

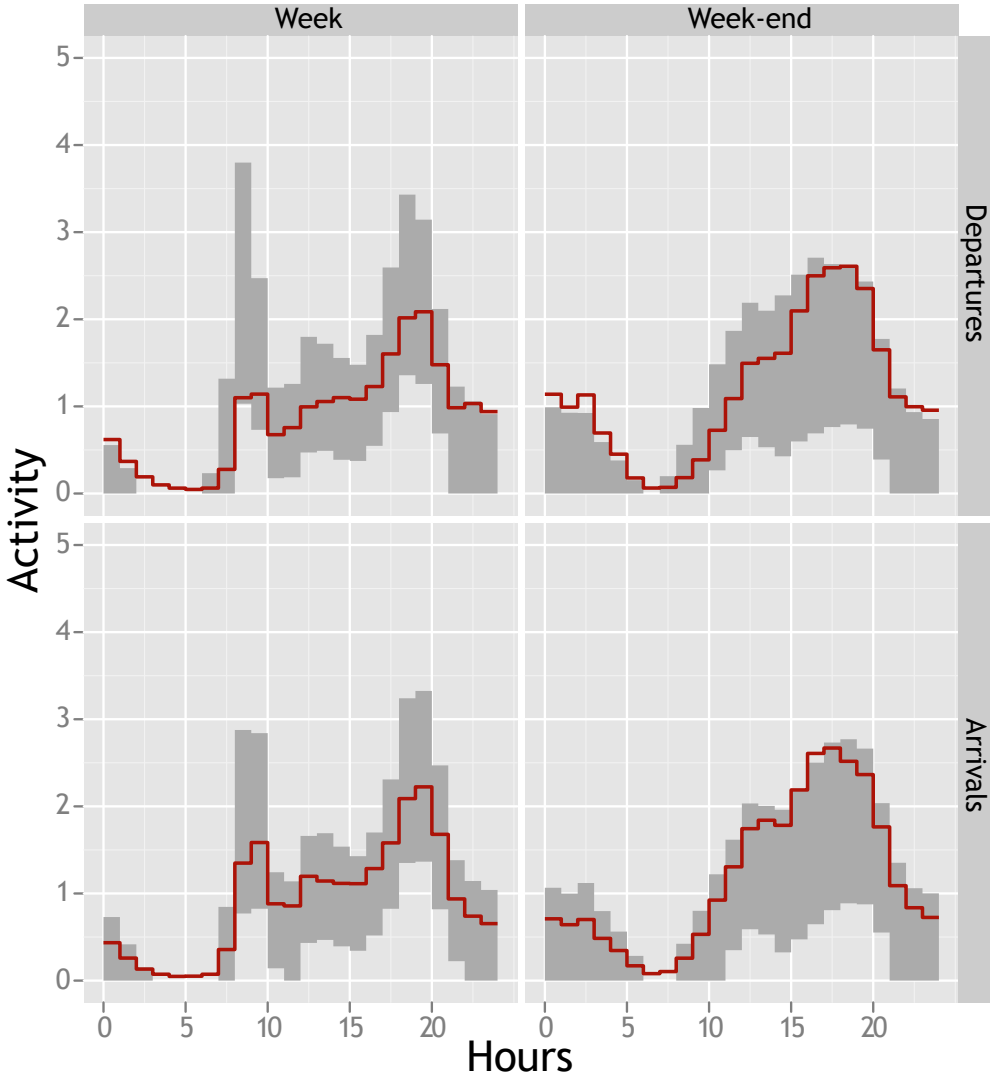
Sorties nocturnes



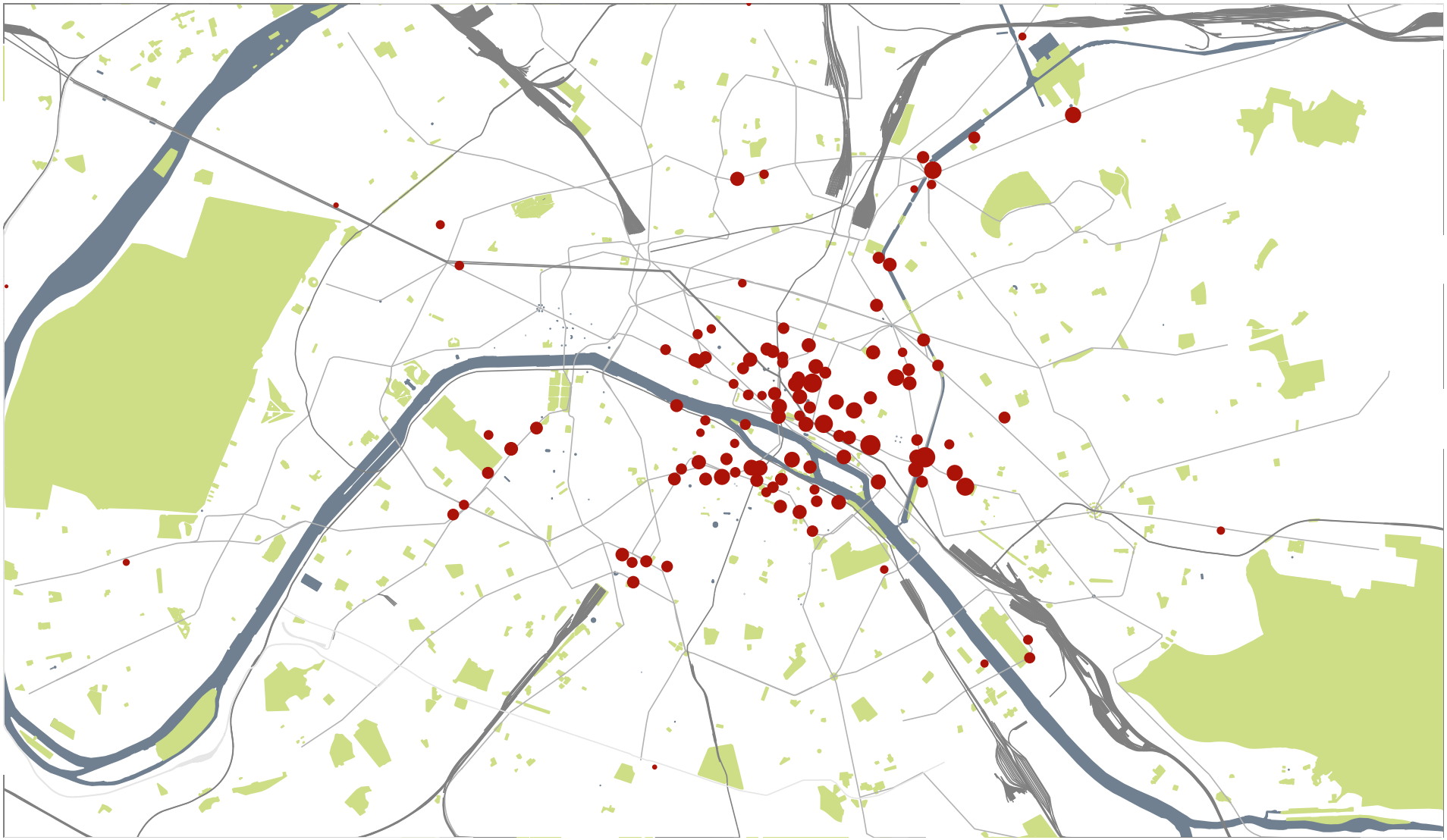
Sorties nocturnes



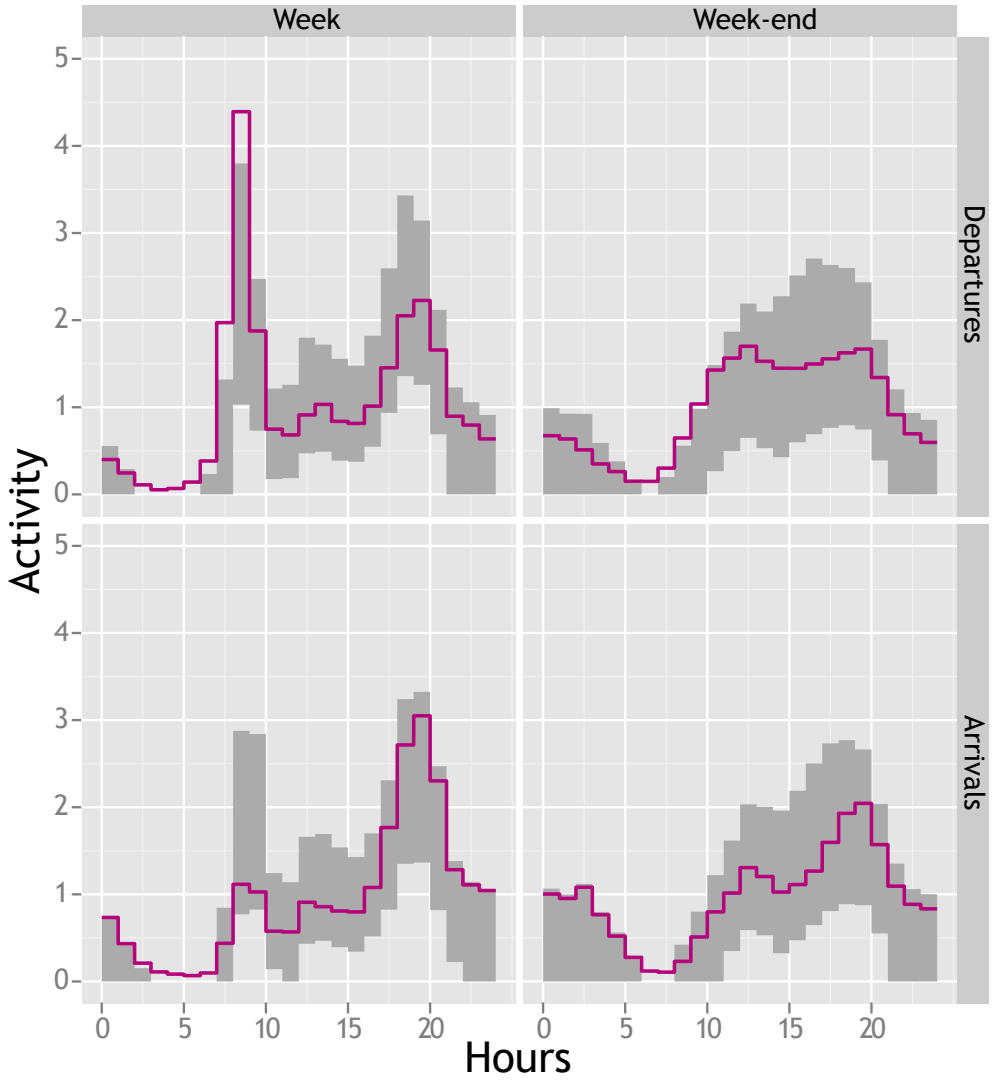
Sorties nocturnes et week-end

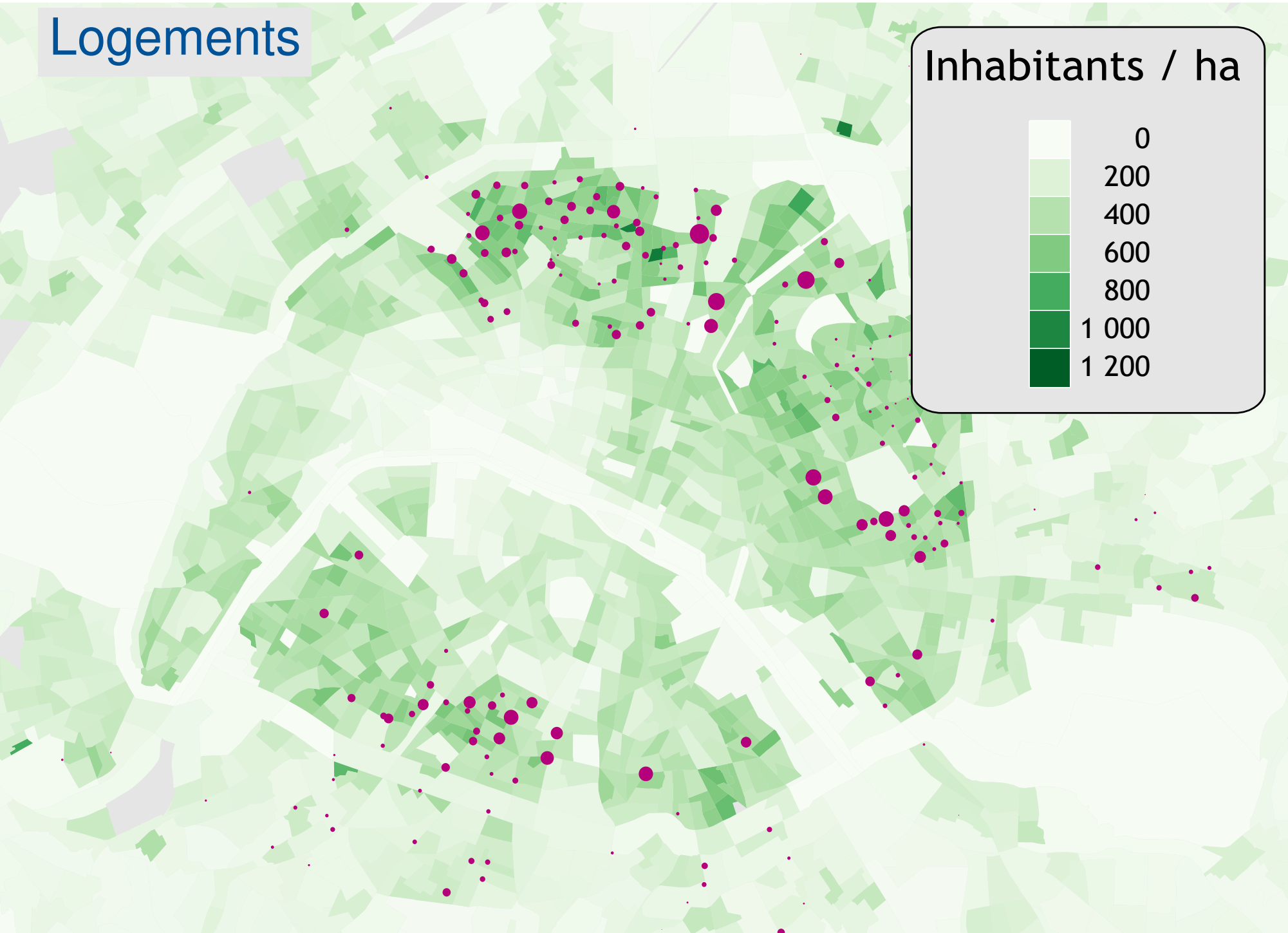


Sorties nocturnes et week-end

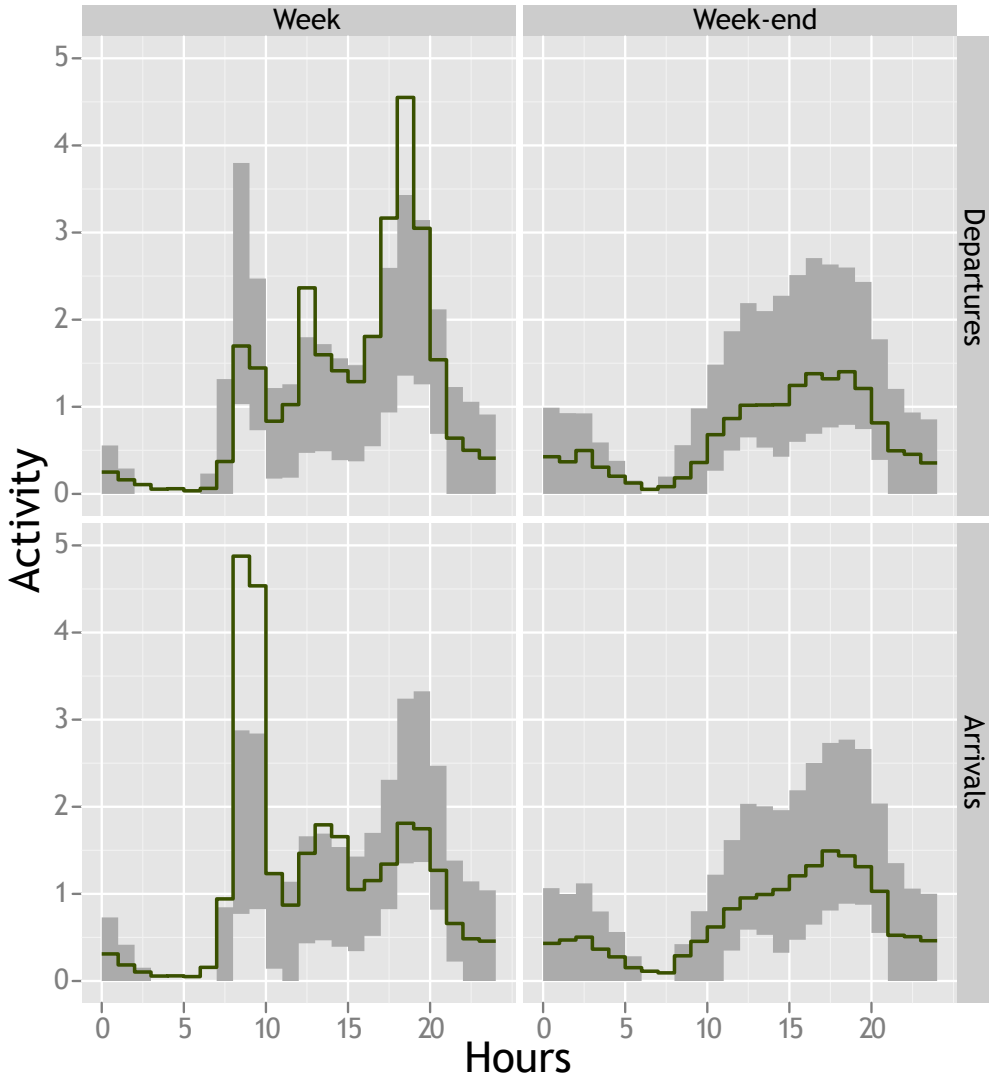


Logements

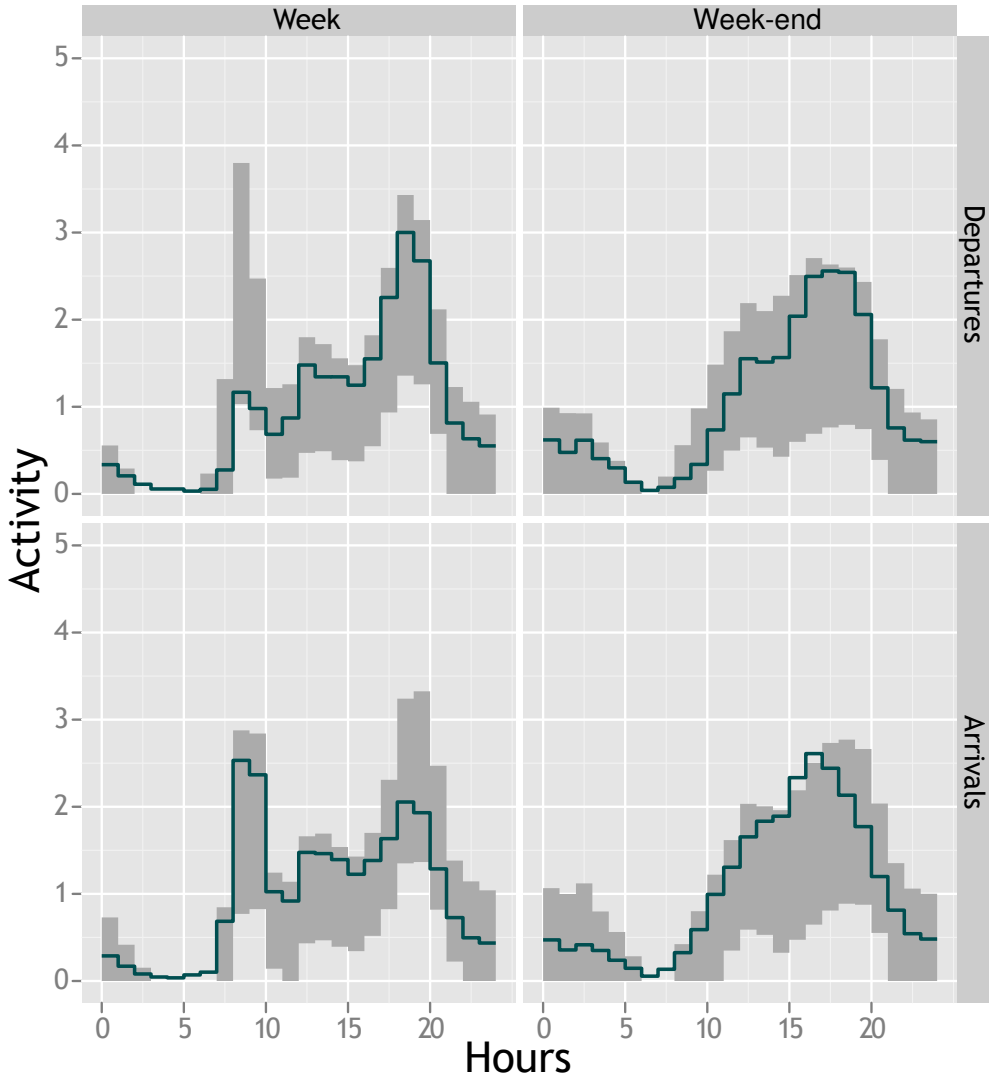

Week
Week-end
Departures
Arrivals
Activity
Hours
0
1
2
3
4
5
0
5
10
15
20

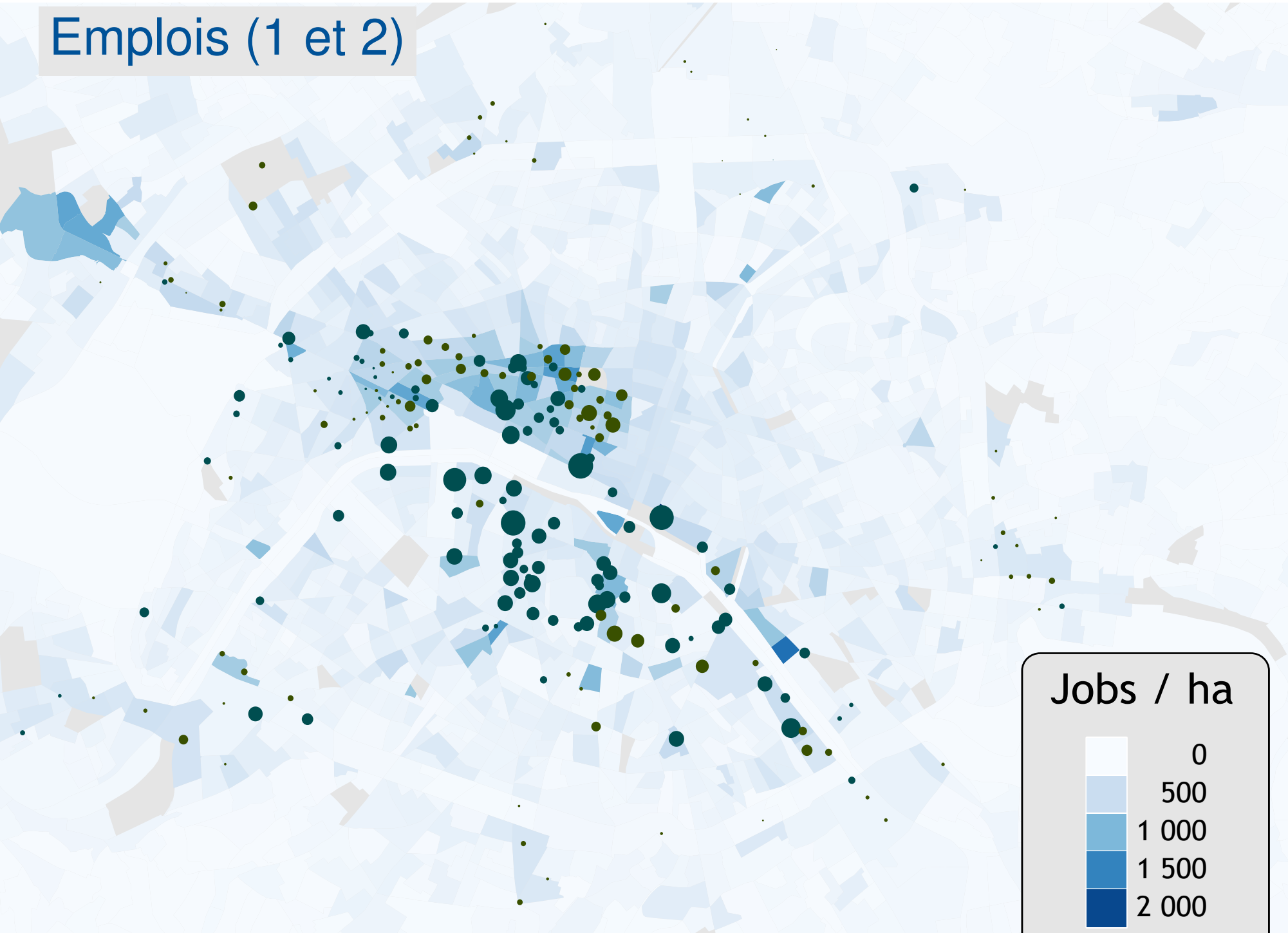

Logements
Inhabitants / ha
0
200
400
600
800
1 000
1 200

Emplois(1)

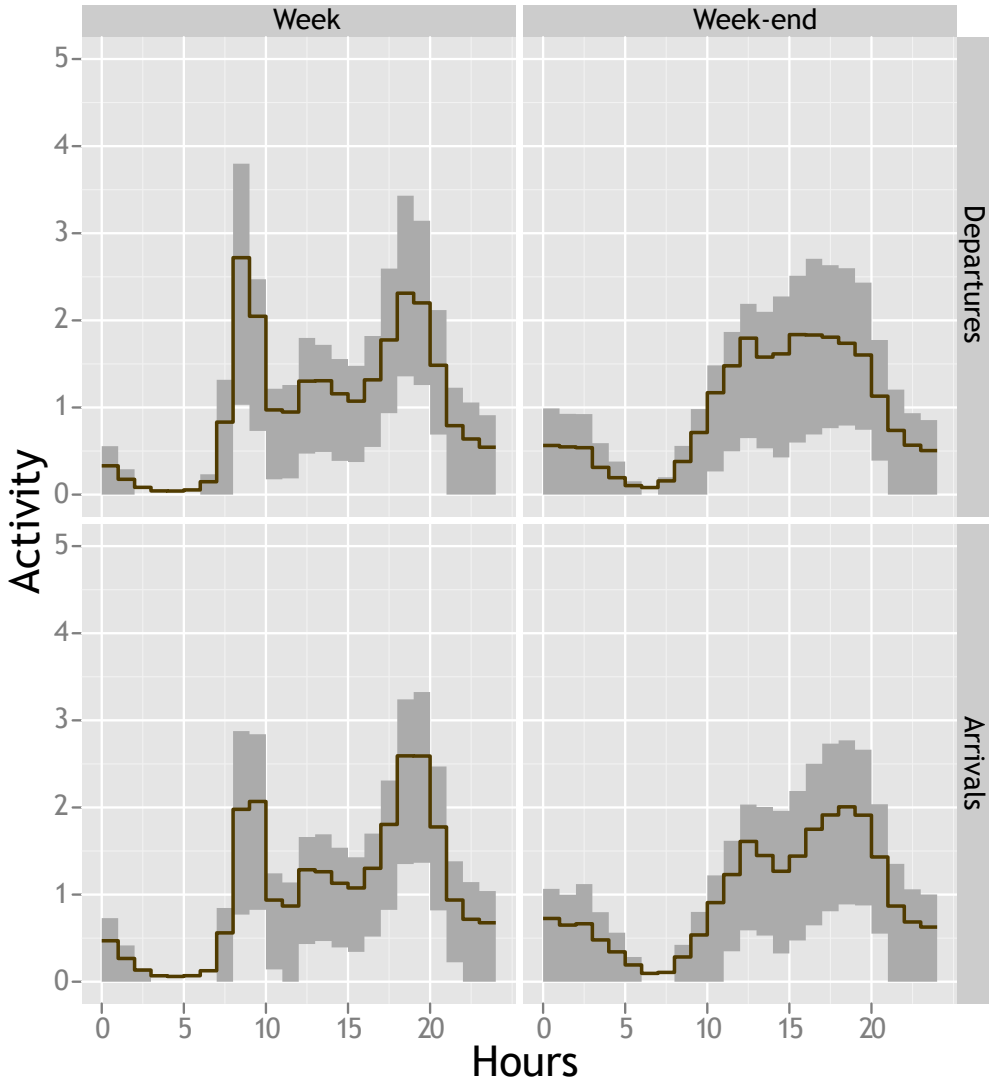

Week
Week-end
Departures
Arrivals
Activity
Hours
0
1
2
3
4
5
0
5
10
15
20

Emplois (2)


Week
Week-end
Departures
Arrivals
Activity
Hours
0
1
2
3
4
5
0
5
10
15
20


Emplois (1 et 2)
Jobs / ha
0
500
1 000
1 500
2 000

Usage mixte


Week
Week-end
Departures
Arrivals
Activity
Hours
0
1
2
3
4
5
0
5
10
15
20

Usage mixte

Croisement données population-emplois-services

	hab/ha	emp/ha	serv/ha	com/ha
	162	237	4.2	3.7
Sorties (1)	367	189	6.3	4.4
Sorties (2)	261	322	7.7	6.9
Parcs	172	90	2	1.7
Gares	209	206	2.4	1.8
Logements	375	108	3.8	2.7
Emplois(1)	138	409	4.5	2.8
Emplois(2)	157	456	5.7	5.6
Moyennes	301	163	3.8	2.8

TABLE 1: Comparaison des moyennes de densités de population, d'emplois, de services et de commerces pour les différentes groupes de stations.

Conclusions sur le clustering de stations

Résultats

- Des stations bien différenciées en termes d'usage
- Interprétation aisée des groupes de stations
- Densités de population, d'emplois et d'équipements explicatives des groupes de stations
- Profils temporels des clusters interprétables et informatifs

Exemple 3 :

Clustering de matrices O-D dynamiques

Clustering de matrices O-D dynamiques

Objectifs

- Représentation des données : matrices OD dynamiques
  $\Rightarrow$ Recherche de stationnarités et de points de changement dans la dynamique des matrice OD
- Modèle utilisé “Latent Dirichlet Allocation”

Résultats

- Segmentation temporelle et cycles
- Distribution spatiale des flux (OD de référence) / segment
  $\Rightarrow$ Permet de caractériser le déséquilibre du réseau / segment

Représentation synthétique de la dynamique du réseau

LDA, pour matrice O/D dynamique

- pour chaque activités $a \in \{1, \ldots, N_a\}$, tirer un générateur de matrice O/D :

$$\Lambda_a \sim \mathcal{D}(\beta) \tag{28}$$

- pour chaque sac de déplacement $t \in \{1, \ldots, N_t\}$:
  1. tirer les proportions des activités :

$$\pi_t \sim \mathcal{D}(\alpha) \tag{29}$$

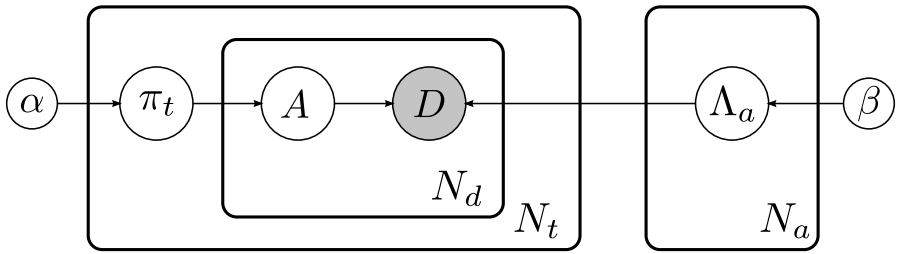
  2. pour chaque déplacement de la tranche horaire t :
     - tirer l'activité A du déplacement :

$$A \sim \mathcal{M}(1, \pi_t) \tag{30}$$

     - tirer un déplacement D (couple O/D) conditionellement à l'activité :

$$D \sim \mathcal{M}(1, \Lambda_A) \tag{31}$$

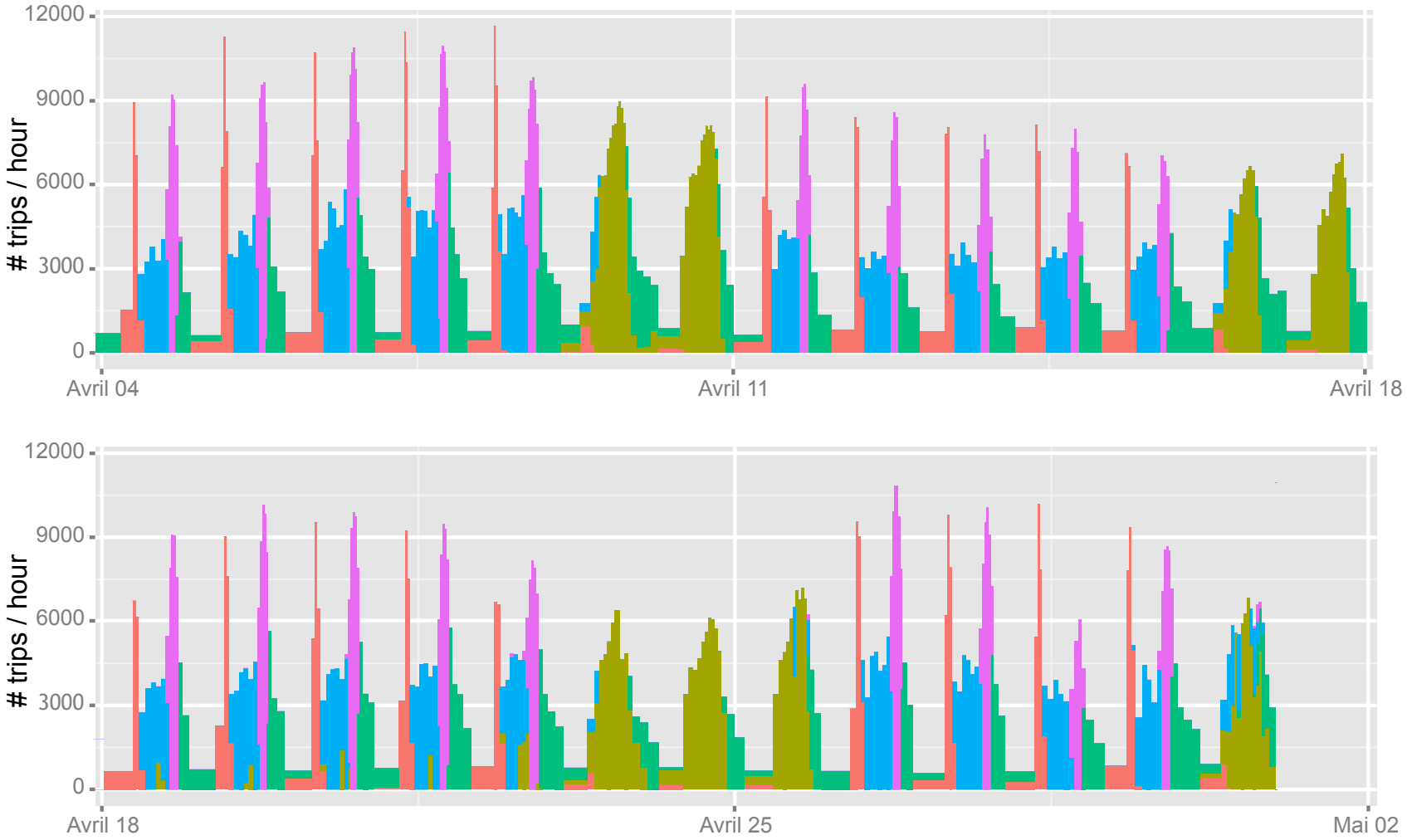
LDA, pour l'analyse de matrice O/D dynamiques



Données d'entrées et pertinence des sorties

- Données d'entrées : comptage d'OD observées durant des fenêtres temporelles
- Sorties :
  1. π_t : proportions des activités latentes pour chacune des fenêtres
     $\Rightarrow$ segmentation temporelle (plage de stationnarité)
  2. Λ_a : ensemble de matrices O/D type pour chaque activité latente
     $\Rightarrow$ caractérise le comportement du système

Interprétation temporelle



Interprétation temporelle

Remarques

- La cyclostationnarité est clairement visible
- Faible mélange entre les différentes activités latentes
- Interprétation temporelle des 5 groupes obtenus : Domicile $\leftrightarrow$ Travail, Déjeuner, Travail $\leftrightarrow$ Domicile, Loisirs nocturnes, Loisirs

Interprétation spatiale : déséquilibre du réseau

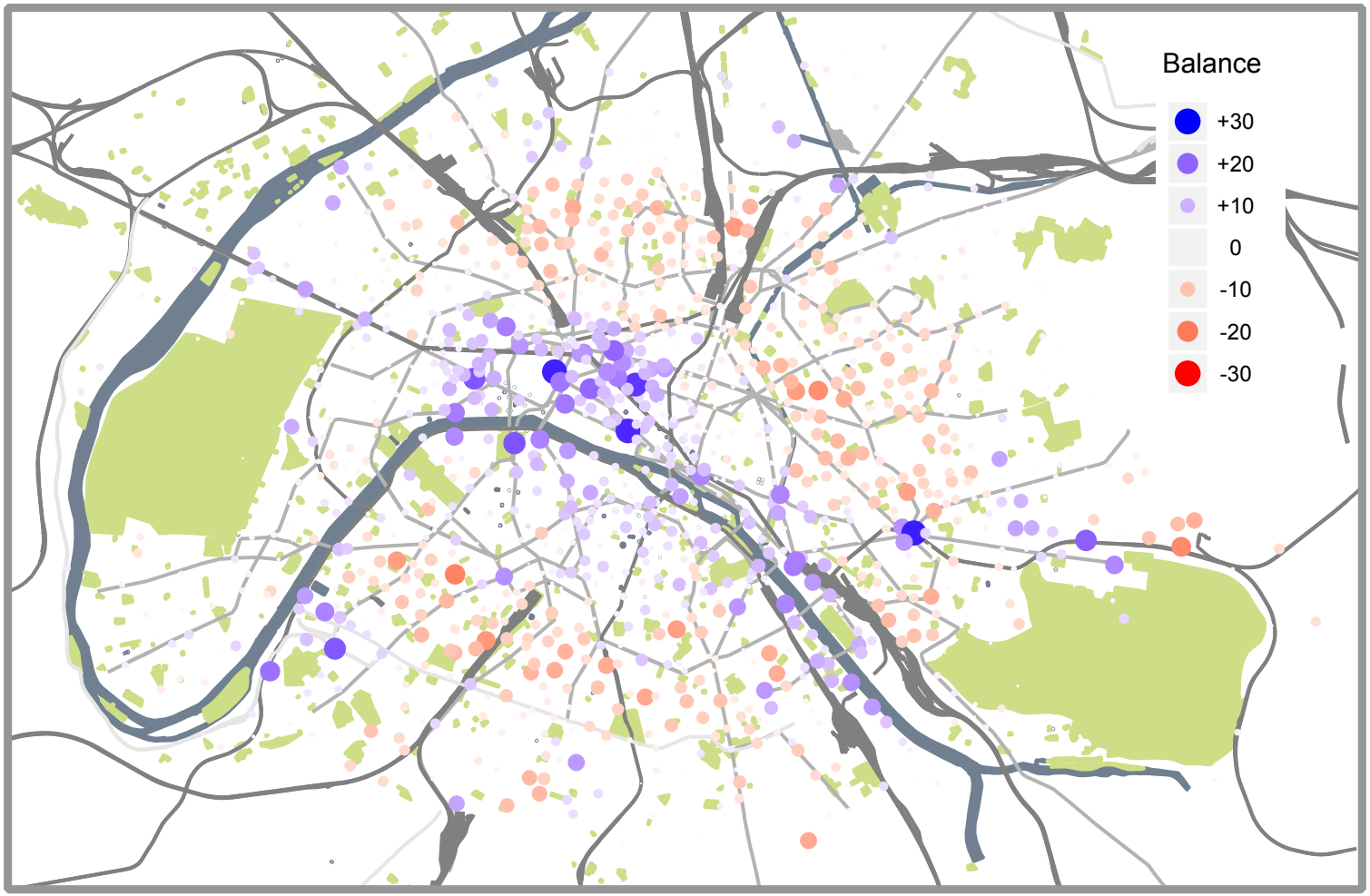


FIGURE 13: Activité latente "Domicile→Travail", déséquilibre du réseau pour $N_{dep} = 10\,000$

"Travail→Domicile"

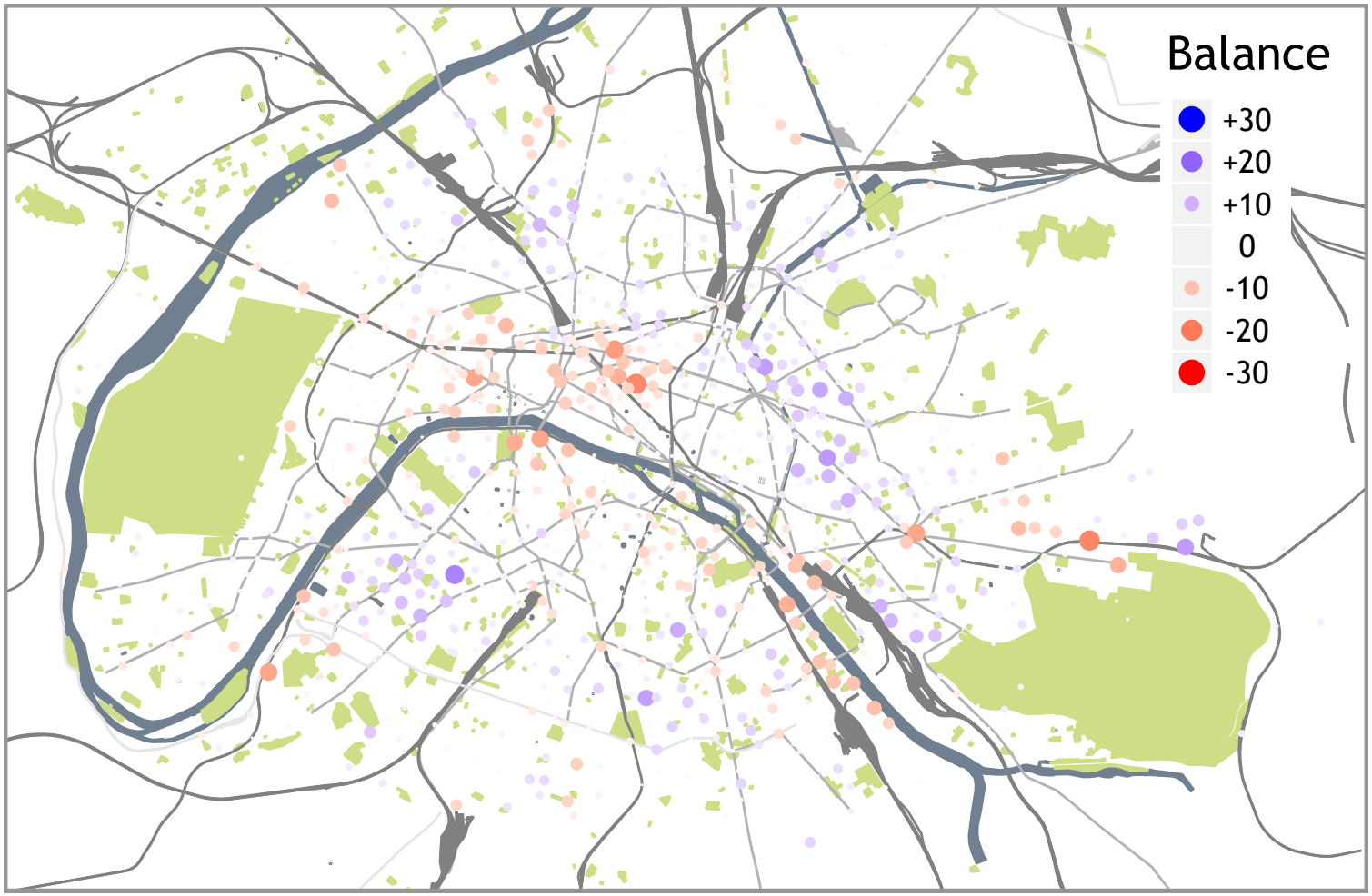


FIGURE 14: Activité latente "Travail→Domicile", déséquilibre du réseau pour $N_{dep} = 10\,000$

"Déjeuner"

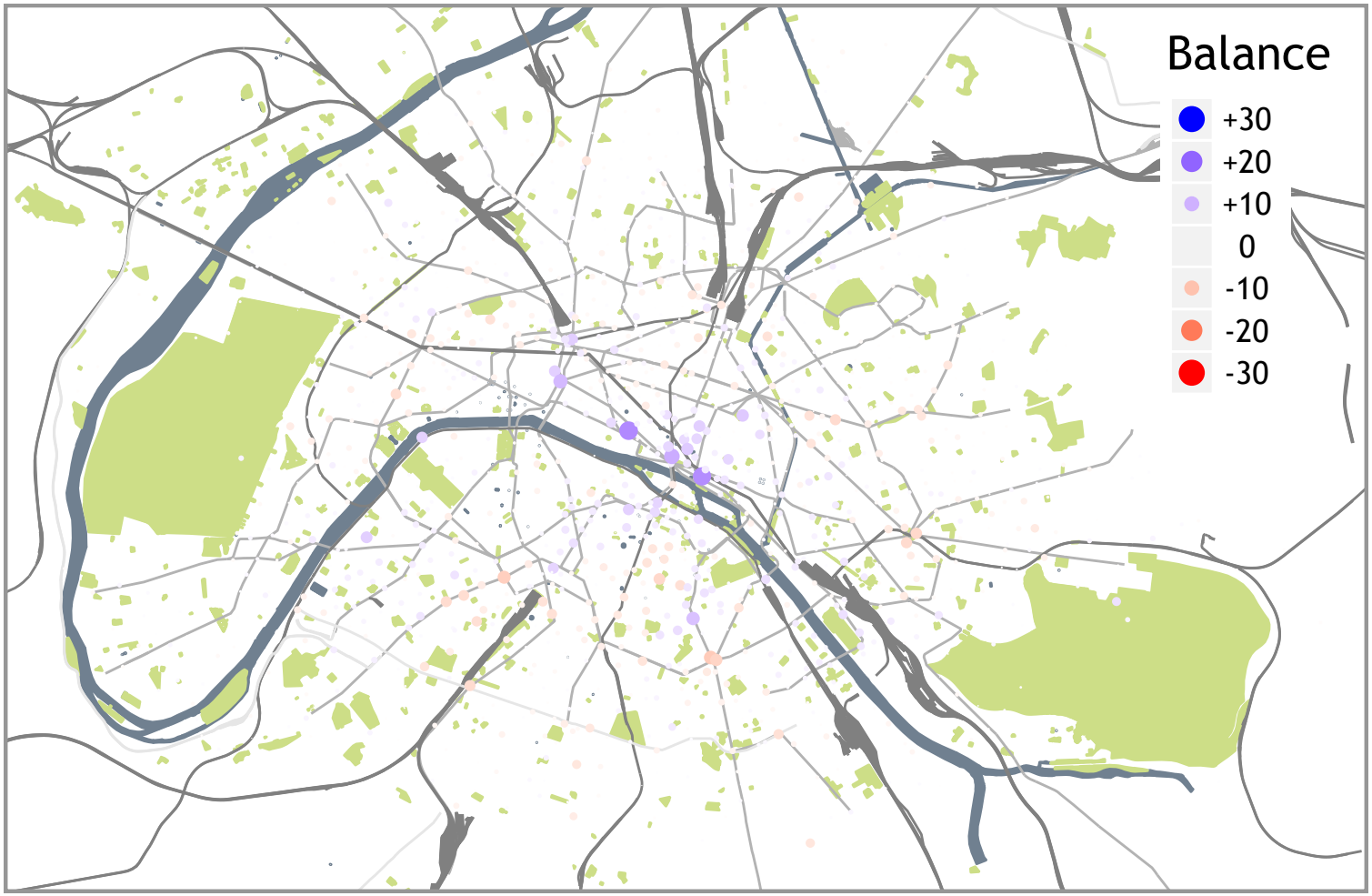


FIGURE 15: Activité latente "Déjeuner", déséquilibre du réseau pour $N_{dep} = 10\,000$

"Loisirs nocturnes"

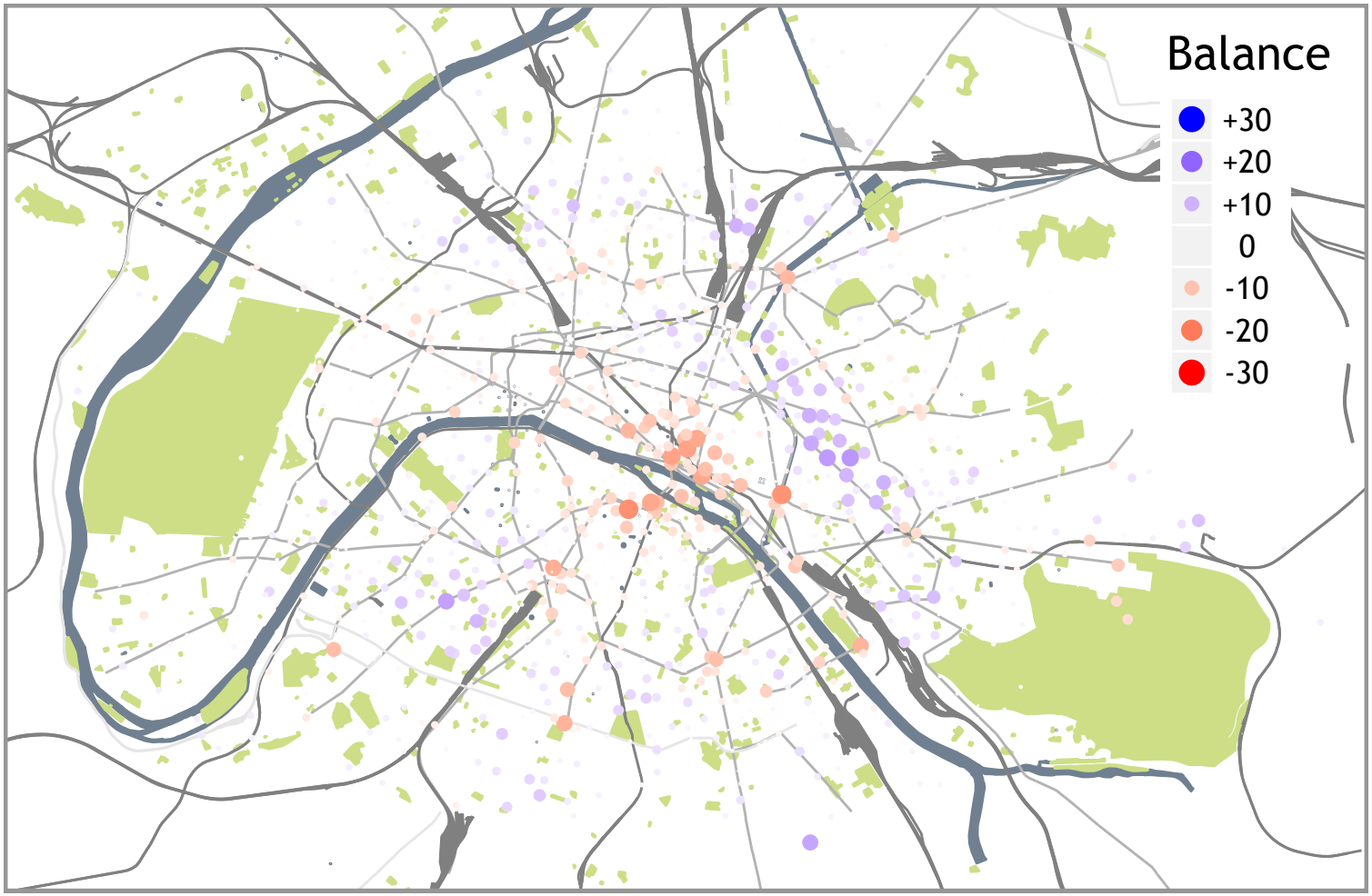


FIGURE 16: Activité latente "Loisirs nocturne", déséquilibre du réseau pour $N_{dep} = 10\,000$

"Loisirs"

FIGURE 17: Activité latente "Loisirs", déséquilibre du réseau pour $N_{dep} = 10\,000$

"Loisirs", stations avec un fort flux entrant

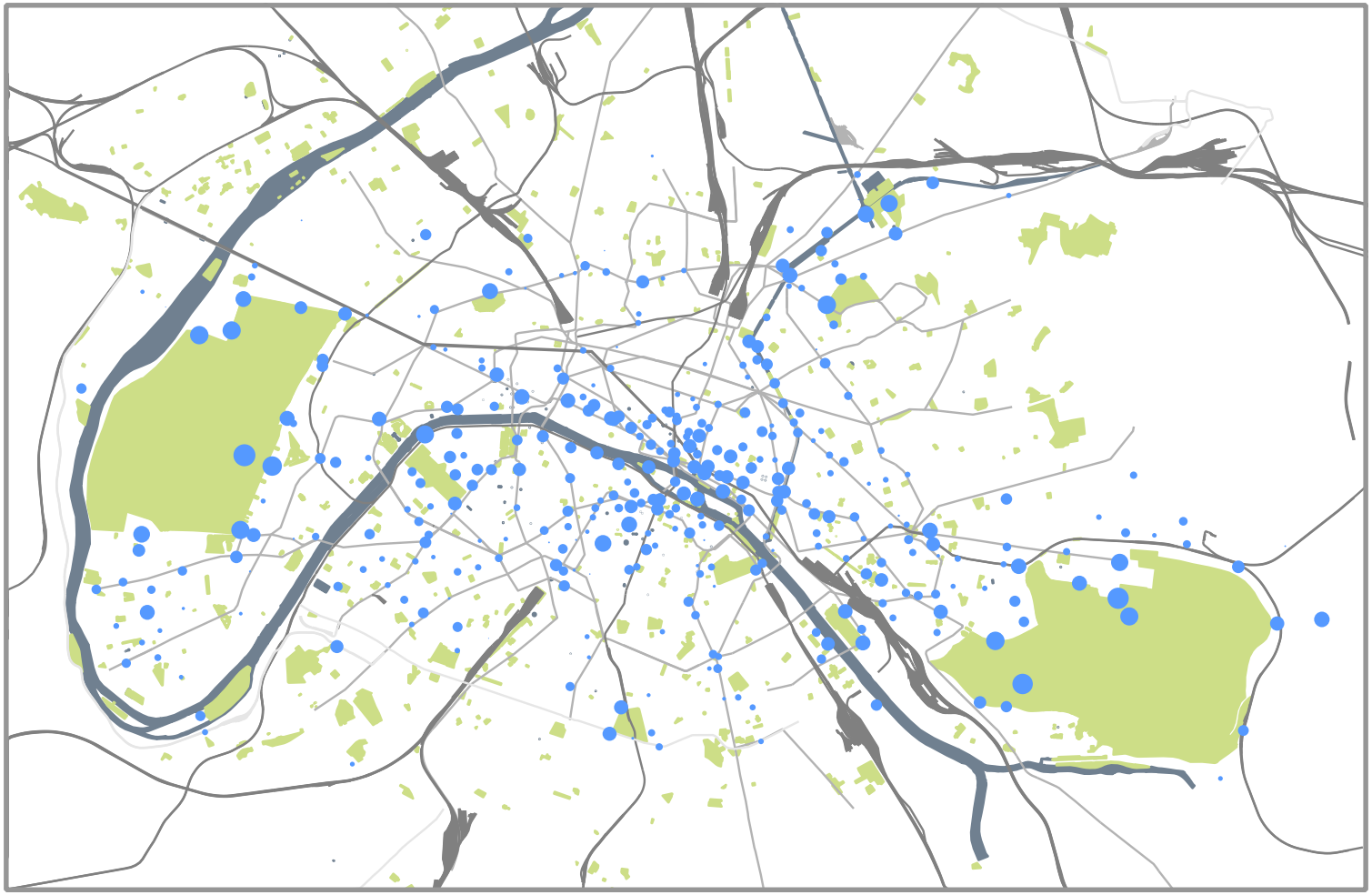


FIGURE 18: Stations incoming specificity

"Loisirs", stations avec un fort flux sortant

FIGURE 19: Stations outgoing specificity

Résultats

- Meilleure compréhension de la dynamique du réseau
- Modèle synthétique de la dynamique
- Mise en évidence des cycles

Limites

- Prise en compte uniquement de la demande satisfaite
- Pas de liens explicites avec les données socio-économiques et géographiques

Travaux actuels & futurs

Aspects intermodaux, données billetique

Synergie entre VLS et TC

- Objectif : comprendre et mesurer l'intermodalité grâce à des données de billettique

Projet PREDIT Mobilletic (cas d'étude sur la ville de Rennes)
Partenaires : GRETTIA, LVMT, CEREMA - CETE NP, Keolis Rennes

Autour des modèles a variables latentes

- LDA avec covariables socio-économiques (thèse A. Randriamanamihaga)
- Clustering de profil d'utilisateurs "mixture of unigrame".
- Prediction : Régression beta-poisson (phénomènes de saturation / stations)

Merci pour votre attention !

etienne.come@ifsttar.fr
mobilletic.ifsttar.fr
vlsstats.ifsttar.fr
www.comeetie.fr
@comeetie

Ifsttar
Centre de Marne-la-Vallée
Batiment le “Bienvenue”
14-20 Bd Newton Cité Descartes, Champs sur Marne
F-77447 Marne la Vallée Cedex 2

Tél. +33 (0)1 81 66 87 19

References

Govaert, G. and Nadif, M. (2010).

Latent Block Model for Contingency Table.

Communications in Statistics-Theory and Methods 39, 416 – 425.

Rau, A., Celeux, G., Martin-Magniette, M.-L. and Maugis-Rabusseau, C. (2011).

Clustering high-throughput sequencing data with Poisson mixture models.

Rapport de recherche RR-7786 INRIA.

Witten, D., Tibshirani, R., Gu, S., Fire, A. and Lui, W. (2010).

Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumous and matched controls.

BMC Biology 58.