La question du zéro dans le modèle gravitaire

ThéoQuant, Besançon - 19 mai 2017

Hadrien Commenges[1] & **Daniel Commenges**[2]
[1]UMR Géographie-cités, Université Paris Panthéon-Sorbonne
[2]INSERM U1219, Université de Bordeaux

Le bricolage gravitaire

$$T_{ij} = \alpha_i A_i \beta_j E_j d_{ij}^{\theta} \tag{1}$$

- θ (scalaire) : approché par l'aglorithme de Hyman (1969)
- α_i et β_j (vecteurs) : approchés par l'algorithme de Furness (1965)

Valeurs du paramètre de la distance selon le SETRA

Horizon	Travail	Autres Motifs	Secondaires
1970	0,8	2	2
1990	En première approximation, on pourra admettre que les valeurs de l'horizon 1970 peuvent être reconduites pour les études prévisionnelles aux horizons 1990 et 2010.		
2010			

Source : SETRA (1973) Études préliminaires d'infrastructures de transport

Le bricolage gravitaire

Furness et Hyman : algorithmes itératifs appliqués sur la matrice origine-destination dans sa totalité

DESTINATION / ORIGINE	1 2	. . j . .	. . . . n	EMISSIONS ▼
1 2 |				E_1 E_2 |
i		T_{ij}		E_i
| n				| E_n
ATTRACTIONS ▶	A_1 A_2	. A_j .	. . . . A_n	TOTAL

Source : SETRA (1973) Études préliminaires d'infrastructures de transport

La régression gravitaire

$$T_{ij} = A_i^{\alpha} E_j^{\beta} d_{ij}^{\theta} \tag{2}$$

$$log(T_{ij}) = \alpha log(A_i) + \beta log(E_j) + \theta log(d_{ij}) \tag{3}$$

- α, β, θ estimés par régression linéaire OLS (années 1970)
- α, β, θ estimés par régression de Poisson (Flowerdew & Aitkin 1982)

La question du zéro

Causes de la disparition du zéro

- Dans certains cas d'étude (graphes complets) il n'y a pas de zéros
- La régression avec transformation en log n'admet pas de zéros
- Les fournisseurs de données ne fournissent pas les zéros

Problèmes posés par la présence de zéros

- Quand on prend en compte les zéros, il y en a trop (calcul numérique)
- Quand on prend en compte les zéros, il y en a trop (surdispersion)
- Quand il y a trop de zéros, il est difficile de les filtrer (*zero-inflated*)

La question du zéro

Résultat des courses

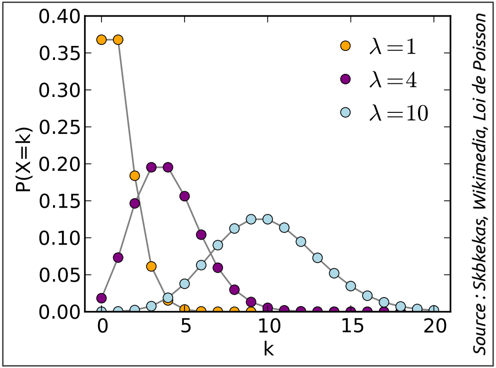
- Pas de réflexion sur le statut du zéro : **absence d'information** ou **absence d'interaction**
- On travaille sans les zéros sans prendre en compte ce filtre sur la variable à expliquer

$\rightarrow$ la régression de Poisson (sous différentes formes) s'est imposée or la probabilité d'obtenir des zéros dans une telle distribution est non nulle

Distribution statistique de Poisson

Un seul paramètre λ, à la fois la moyenne et la variance de la distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Régression de Poisson avec troncature

La régression de Poisson considère que Y_i suit une distribution conditionnelle de Poisson dont la moyenne μ_i s'écrit :

$$\log \mu_i = \beta_0 + \beta_1 X_{i1} \ldots + \beta_p X_{ip} = X_i \beta, \tag{4}$$

$$\mathcal{P}(Y = y) = \frac{\mu^y e^{-\mu}}{y!} \tag{5}$$

La log-vraisemblance pour i s'écrit :

$$L_i(\beta) = Y_i X_i \beta - \exp(X_i \beta) - \log Y_i! \tag{6}$$

Régression de Poisson avec troncature

Distribution conditionnelle avec troncature :

$$\mathcal{P}(Y = y | Y > 0) = \frac{\mathcal{P}(Y = y)}{\mathcal{P}(Y > 0)} \tag{7}$$

Donc :

$$\mathcal{P}(Y = y | Y > 0) = \frac{\mu^y e^{-\mu}}{y!(1 - e^{-\mu})} \tag{8}$$

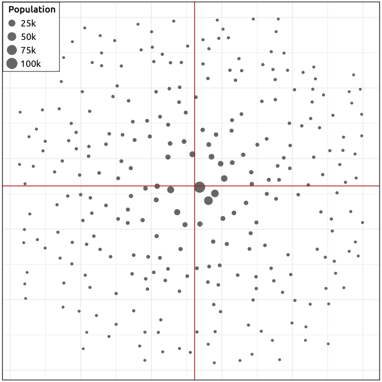
Et donc la vraisemblance conditionnelle :

$$L_{i|Y_i>0}(\beta) = Y_i X_i \beta - \exp(X_i\beta) - \log(Y_i!) - \log(1 - e^{-e^{X_i\beta}}) \tag{9}$$

Simulation (mise en place)

Génération d'une aire urbaine type (taille moyenne & monocentrique) :

- semis de points à peu près régulier et à peu près circulaire
- distribution des populations loi de Zipf : $y_i = kx_i^{-\alpha}$
- relation population/distance au centre selon loi de Clark : $y_i = ke^{-\beta x_i}$



Simulation (mise en place)

Génération des flux :

- Génération des μ_i avec un modèle gravitaire classique $\mu_i = A_i^{\alpha} E_j^{\beta} d_{ij}^{\theta}$ où $\alpha = 1$, $\beta = 1$ et $\theta = -2$
- Génération de variables aléatoires de Poisson de moyenne μ_i
- Estimation des cœfficients (moyenne et écart-type) sur 100 réplications

Comparaison de modèles pour l'estimation :

1. Régression gaussienne sans zéros avec transformation en log
2. Régression poissonienne avec zéros (approche rêvée)
3. Régression poissonienne sans zéros et sans prise en compte de la troncature (approche classique)
4. Régression poissonienne sans zéros avec prise en compte de la troncature (approche proposée)

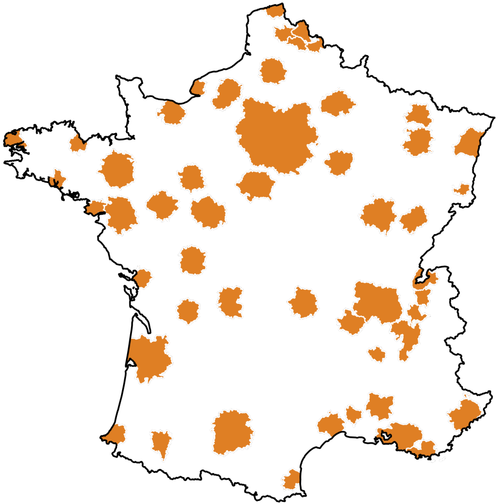
Simulation (résultats)

Rappel sur le modèle : $\mu_i = A_i^{\alpha} E_j^{\beta} d_{ij}^{\theta}$ où $\alpha = 1$, $\beta = 1$ et $\theta = -2$

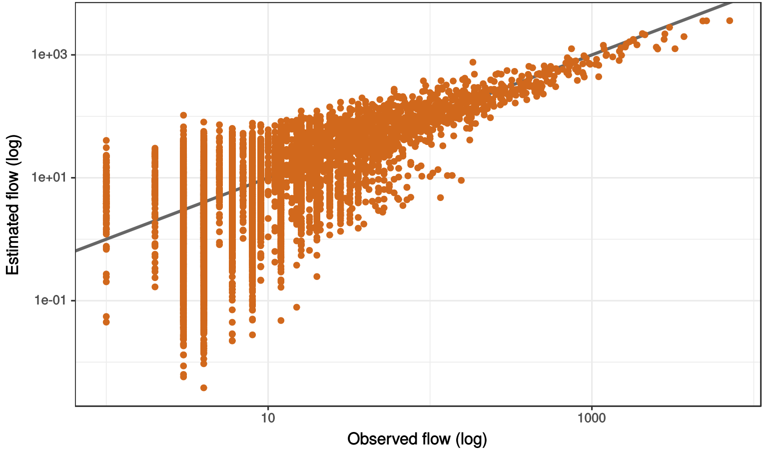
	alpha		beta		theta	
MODEL	MEAN	S.E	MEAN	S.E	MEAN	S.E
(1) Gaussian with log transform	0.6111	0.0034	0.6106	0.0032	-1.1524	0.0072
(2) Poisson with zeros	1.0001	0.0010	0.9998	0.0011	-2.0000	0.0017
(3) Poisson without zeros	0.9847	0.0011	0.9845	0.0011	-1.9712	0.0024
(4) Poisson with conditional likelihood	1.0001	0.0011	0.9999	0.0013	-1.9999	0.0022

Application empirique (mise en place)

- **Données :** navettes domicile-travail RP2010
- **Espace :** 50 premières aires urbaines françaises (35,5 M hab.)

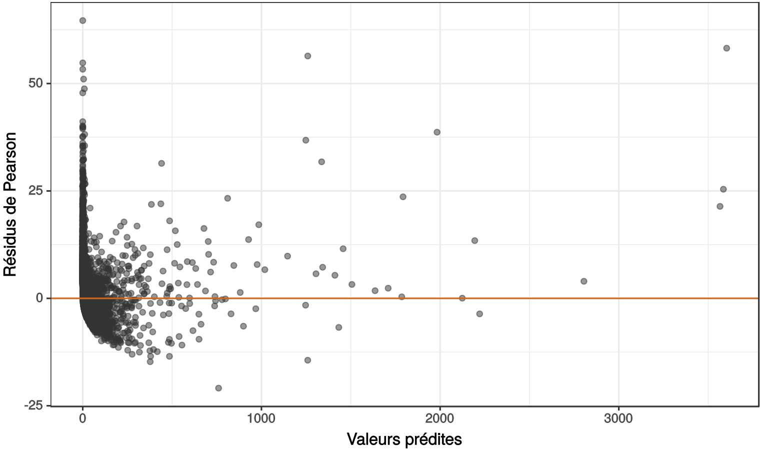


Application empirique (Toulouse)



$\rightarrow$ Cœf. de dispersion (χ^2 rapporté aux d.f.) $=$ 32

Application empirique (Toulouse)



Application empirique (surdispersion)

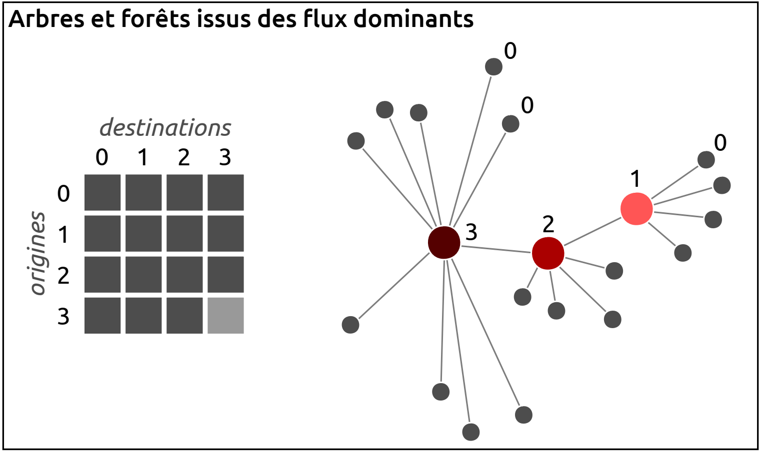
Premières conclusions :

- Prise en compte de la troncature inutile
- Fonction-lien loin d'être poissonienne

Causes possibles de la surdispersion :

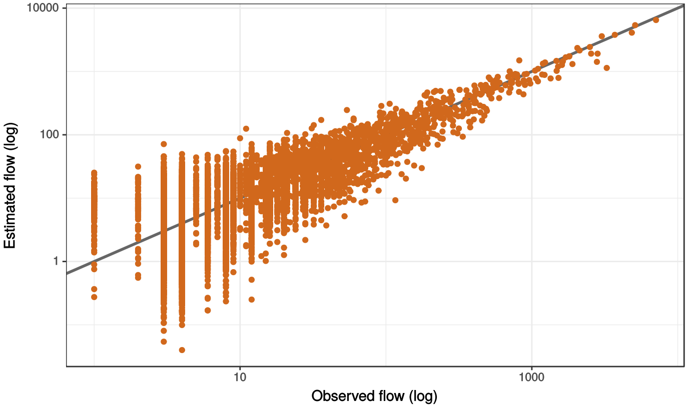
- violations fondamentales (indépendance des observations)
- modèle incomplet (manque des variables explicatives)
- modèle incomplet (manque des termes d'interaction)
- valeurs aberrantes
- fonction-lien mal spécifiée

Typologie des flux



Source : Berroir, Mathian, Saint-Julien, Sanders (2004) Mobilités et polarisations

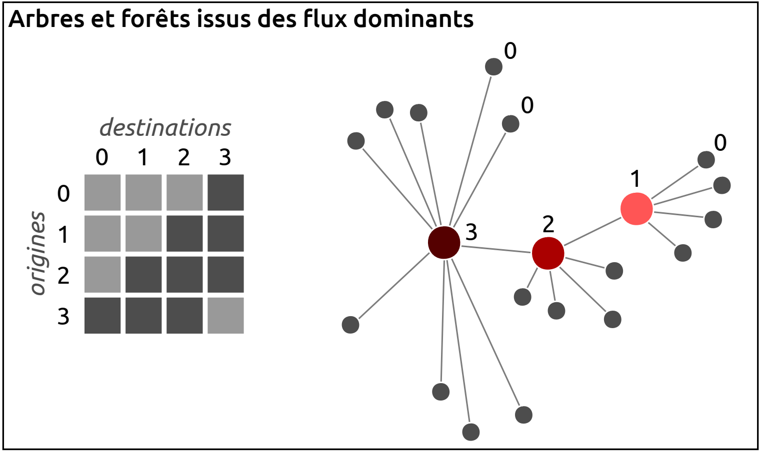
Typologie des flux



Introduction de la typologie de flux comme régresseur :

- Meilleure spécification : cœf. de dispersion = 2
- Modèle illisible : 15 termes d'interaction

Typologie des flux



Source : Berroir, Mathian, Saint-Julien, Sanders (2004) Mobilités et polarisations

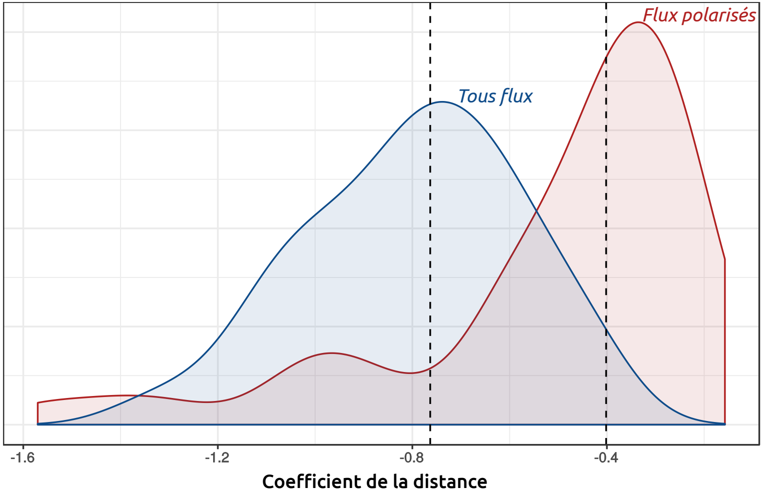
Modèle gravitaire partiel

Distribution des flux (ex. Toulouse)

CLASS	NB	SUM	PCTNB	PCTSUM
0	195085	0	95.07	0.00
[0,5)	4883	16998	2.38	3.77
[5,50)	4361	67052	2.13	14.88
[50,500)	767	108317	0.37	24.04
[500,5000)	110	127499	0.05	28.30
[5000,max]	3	130700	0.00	29.01

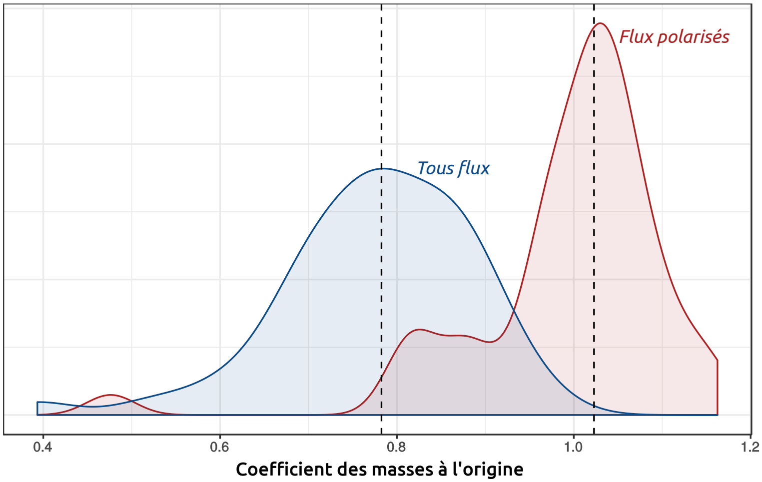
Estimation des coefficients

Distribution des cœfficients de la distance (50 AU)



Estimation des coefficients

Distribution des cœfficients des masses à l'origine (50 AU)



Conclusion

What was the good of confession when you loved the result of your crime ?

And when we love our sin then we are damned indeed.

Graham Greene, The power and the glory