A Finite Mixture Approach to Cost-Function Estimation Economies of Density in Home-Care Services:

Solen Croiset[a], Robert Gary-Bobo[a,b,c]

[a]Université Paris I Panthéon-Sorbonne, CES,
[b]CREST, [c]CEPR.

TSE, Toulouse, 3rd Health Conference, 18 June 2025

Introduction

- The present research is devoted to the micro-econometric study of the costs of providing *home care services* to the elderly (or the disabled) (*services d'aide à domicile*)
- This is both an important and growing sector and a highly regulated, highly subsidized industry (in France).
- We more precisely present here an econometric analysis of *transportation costs* in the home care sector.
- Using a unique dataset, we show the existence of *economies of density* (related to transportation) in this activity.
- *Economies of density* are a particular form of the *economies of scale* (a property of the production function).

Economies of Density : an Important Phenomenon

- Where do we find economies of density ? Cable television is a well-known example. The cost of the cable *per customer* clearly decreases with the *density of population* in the city served by the cable TV provider.
- Other examples : postal services ; garbage collection ; package deliveries (Amazon, UPS, ...)
- Economies of density are tied to geographical space.
- Behind the economic and statistical facts, there is an *Operations Research* problem : Home Healthcare Routing and Scheduling of Multiple Caregivers... (is related to the *Traveling Salesman Problem* in Operations Research).

Organization of Home Care Services

- How do these Home Care Services work ?
- Every morning, in each district, an operator produces a planning of tours and a schedule of visits for employees (caregivers). Employees are mostly women who drive cars across the countryside or suburbs to visit seniors.
- There are various constraints : hours of visit (lunch time, dinner time, etc.) ; clients are accustomed to the carer (they want the same carer to come back every day or week).
- The operator must adjust to different kinds of random shocks : absenteeism and sick leaves of employees ; changes in the demands of clients ; various kinds of accidents.
- So, production is inherently stochastic...

Transportation Costs in Home Care Services

- Kilometers traveled generate costs for two reasons :
  - Hours of work spent travelling between two clients are paid by the employer.
  - Distance traveled is reimbursed to employees by the employer. There is a conventional rate per kilometer (an agreement with unions, *i.e.*, a *convention collective*).
- The employer tries to minimize costs by improvements in routing and scheduling (Operations Management Problem).
- The home care services being heavily subsidized by the government and local governments (*i.e.*, counties or *départements*), public authorities have a stake in the minimization of these costs.
- There are important consequences for Regulation (Natural Monopoly Properties).

Economies of Density : Definition

- On a given territory with given land surface area and given characteristics,...
- ... the density of clients may vary (due, in essence, to variations in market share).
- If kilometers traveled *per hour of service* decrease when the density of clients increases on the territory,...
- ... then, by definition, we have *economies of density* (a purely technological property).
- A territory being given, and all other things being equal (like service quality), an increase in demand for the service is typically always triggered by an increase in the density of clients.

Consequences for Public Policy

- This empirical fact may have important consequences in terms of optimal regulation : the home care services constitute a *natural monopoly*.
- It is then possible to advocate that a system of *franchise bidding*, supervised by the government, would be more efficient than the current system (displaying a mix of competition and regulation that poses many problems).
- The territory of the country should be divided into districts (or constituencies). Each district should be awarded to a single operator for a term, by means of a competitive auction (franchise bidding).

Literature on Economies of Density

- A classic topic in the IO/ Regulation literature.
  - *In Airlines* : Caves, Christensen, Tretheway (1984) ; Braeutigam, Daughety, Turnquist (1984) ; Brueckner Spiller (1994) ;
  - *In Postal Services, Parcel Delivery, Garbage Collection* : Houde, Newberry, Seim (2017) ; Cazals *et al.* (2001) ; Dubin, Navarro (1988) ;
  - *In Railroad Freight Transportation, Container Shipping, International Trade* : Bitzan, Keeler (2007) ; Xu, Itoh (2018) ; Mori, Nishikimi (2002), Behrens, Gaigné, Ottaviano, Thisse (2006).
  - *In Network Industries (electricity, telephone, water utilities)* : Roberts (1986) ; Guldmann (1990) ; Torres, Morrison-Paul (2006) ;
  - *In Personal Service Industries, Chain Stores, Retail Banking* : Morikawa (2011) ; Holmes (2011) ; Aguirregabiria *et al.* (2016) ;

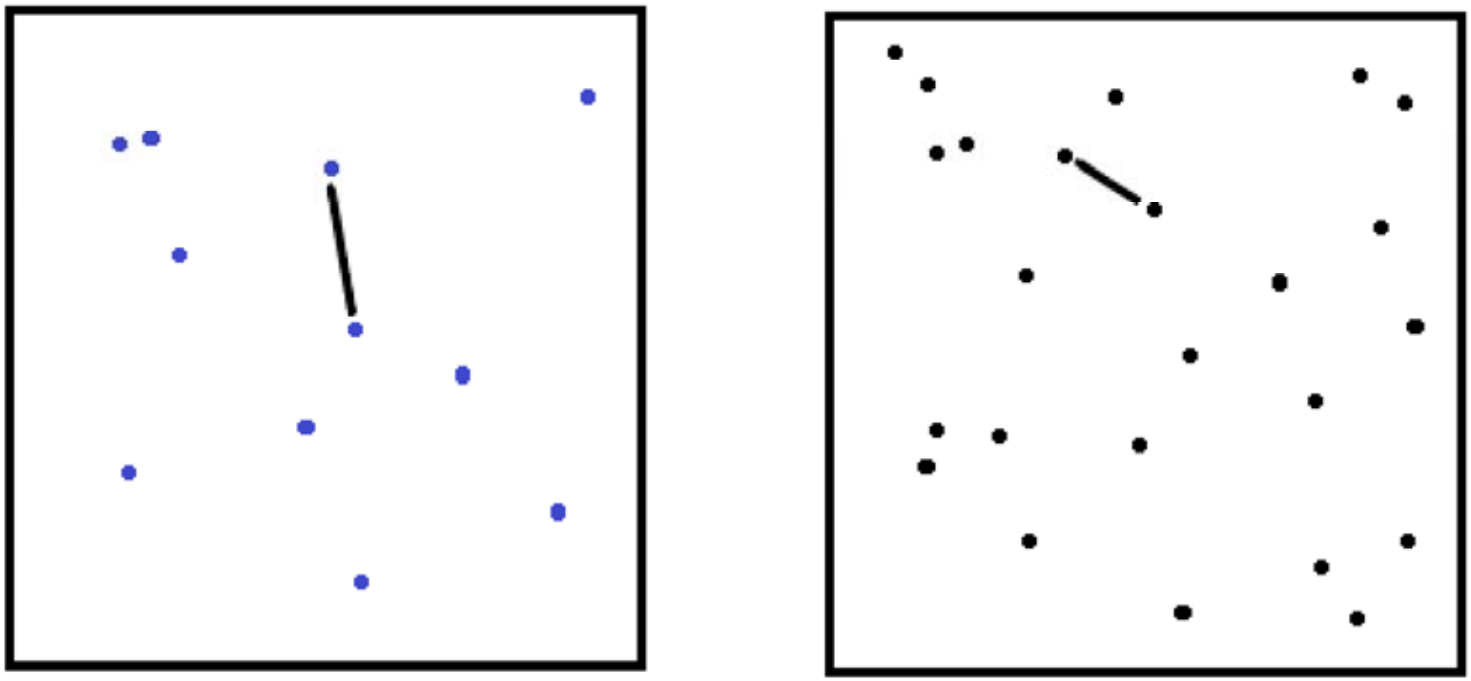
Literature, ctd. Agglomeration Economies.

- More generally, economies of density are related to the theme of *agglomeration economies* (see *e.g.*, the surveys of Combes and Gobillon (2015) ; Rosenthal and Strange (2020)) ;
- The present research shows a pure case of density economies, *i.e.*, the direct result of the interaction of a production technology with geographical space, in the presence of random shocks, without any need for the presence of externalities (on this problematic, see *e.g.*, Ciccone and Hall (1996)).
- To the best of our knowledge, the literature on Home Care Services (and Health Economics) has not focused on economies of density. Papers on the French Home Care Services are due to Gramain and Xing (2012) ; Roquebert and Tenand (2017) but do not address the same questions.

Outline

1. Economies of Density : the Kilometers/Hours (*i.e.*, L/H) ratio.
2. Log-Linear Models
   - Panel-data, Fixed Effects : *Within-group* and First-differences Estimates.
3. Unobserved Heterogeneity and Finite Mixture Model
   - Unobserved Heterogeneity : the Quadratic Model with Latent Types.
   - The U-shaped Average Transportation-Cost Curves.
   - Probabilities of Unobserved Types : Quality of Classification of Employees.
4. Appendix
   - OLS results on Pooled Data
   - Log-linear model at the district level
   - The *Travelling Salesman Problem* and the *Beardwood-Halton-Hammersley* Theorem.
   - Likelihood Function and Maximum Likelihood Estimation.
   - Choice of the Number of Types, Entropy and Information Criteria.

Fig. 1. Distance Between Two Patients and Population Density

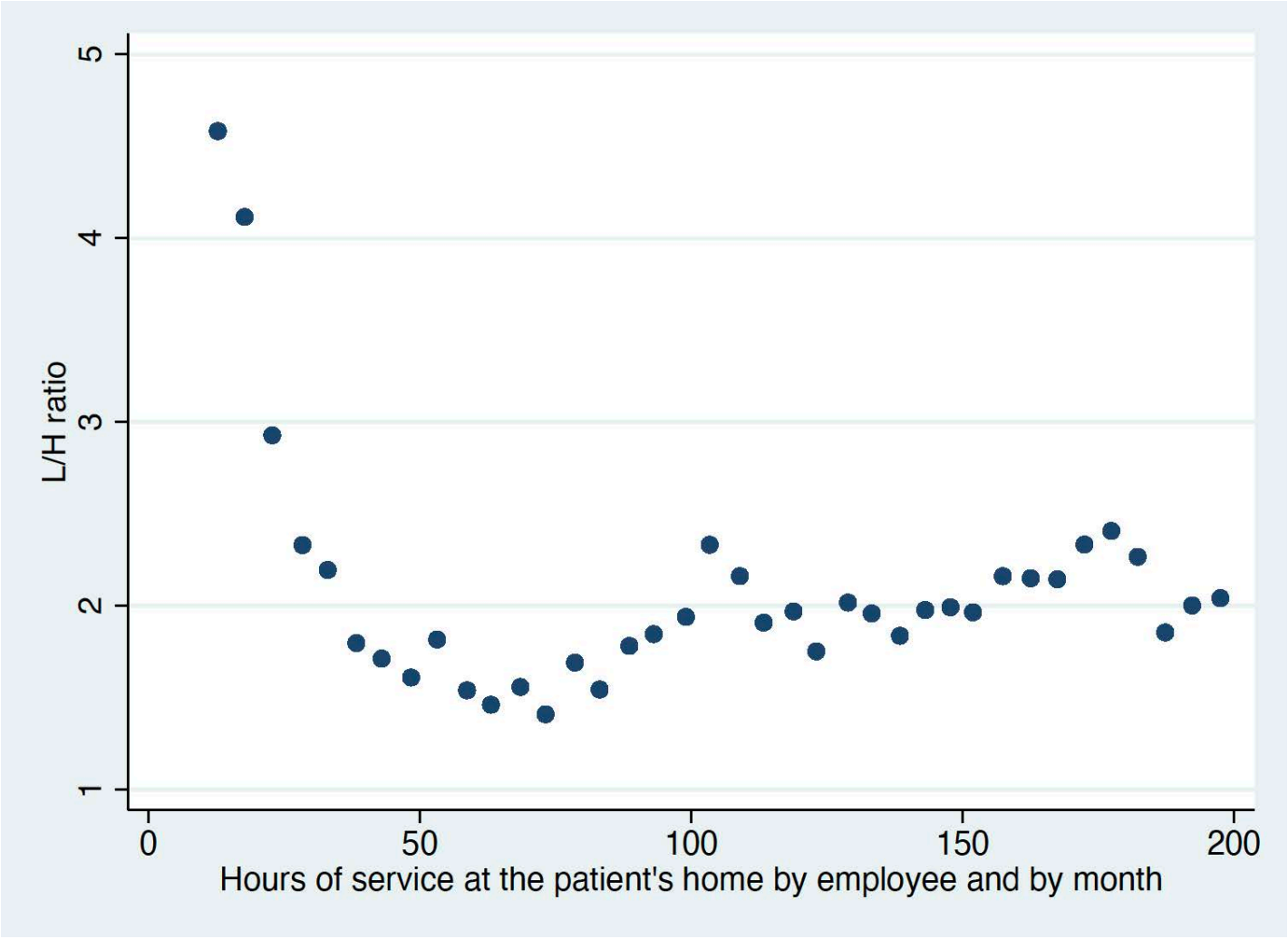


Note : The square depicts a territory. Points are the addresses of patients (seniors needing assistance). In the right-hand square, the number of patients doubled as compared to the left-hand square. Distance to the nearest neighbor decreases by 30%... The area of a disk including the nearest neighbor with (any) probability p is divided by 2 when the number of clients doubles, when points are uniformly distributed.

Ratio of Kilometers Traveled over Hours of Service

- The ratio of kilometers travelled over hours of service at the senior's home allows us to measure the real importance of the economies of density.
- This ratio happens to vary substantially between employees and between the local branches (*i.e.*, *agences*) of the Home Care Service.
- The ratio km/hours is higher in counties (*i.e.*, *départements*) in which the senior population is sparse.
- The ratio km/hours is smaller when the market share of a given home care service is high.

Fig. 2. Km/Hours of Service Ratio at the Employee Level



The Simplest Log-linear Model

- Let L be the traveled distance (in km, per month, of an employee, or in a district).
- Let H be the number of hours of care at home (per month) of an employee (or in a district).
- A is an expression depending on a number of local (district level) factors X, *i.e.*, $A = A(X)$.
- Then we have, approximately,

$$\frac{L}{H} = \frac{A}{H^\gamma}, \quad \text{and} \quad \gamma \geq 0.$$

  Km/hours ratio decreases with the hours of service.
- If $\gamma = 0$, we have "constant returns to scale" $L = AH$.
- If $\gamma = 1$ we have $L = A$, *i.e.*, transportation cost is fixed.
- Economies of density exist as soon as $\gamma > 0$.

Data

- We have a panel with employees indexed by $i = 1, \ldots, N$ observed during τ_i periods $t = 1, \ldots, \tau_i$.
- The panel is unbalanced (some employees appear only in a subset of periods).
- Each employee i appears only in one district (*i.e.*, *secteur*) s. Districts partition the set of employees.
- The panel has the following characteristics
  - Number of counties (*i.e.*, *départements*) : 16.
  - Number of branches (*i.e.*, *agences*) : 53.
  - Number of districts (*i.e.*, *secteurs*) : 98.
  - Number of employees : 3688.
  - Mean number of observed months, *i.e.*, mean value of $\tau_i \simeq 29$.
  - Number of observations (*i.e.*, number of (i, t)) : 56 830.
- All observations come from the payrolls of the *Avec* nextwork of Home Care services. A network of nonprofit organizations (*i.e.*, *associations*).

Econometrics 101 : the Simplest Model

- We denote $\ell_{it} = \log(L_{it})$ et $h_{it} = \log(H_{it})$.
- Taking logarithms we have,

$$\ell_{it} = \alpha + \beta h_{it} + X_{it}\delta + \lambda_s + v_{it},$$

- where $i =$ employee, $t =$ month and λ_s is a district fixed effect. Districts are indexed by $s = 1, \ldots, S$. We can add controls X_{it}, and v_{it} is a random error term.
- **There are economies of density iff** $\beta < 1$, since

$$\beta = 1 - \gamma.$$

- If we run this regression, we find, in essence, $\hat{\beta}_{OLS} \simeq 1$: "constant returns".

Econometrics 2 : Group Synergies ; Model A

- A_{st} is the subset of agents active in district s during month t.
- We aggregate at the district level. Define the average hours per month in district s during month t,

$$\bar{h}_{st} = \frac{1}{n_{st}} \sum_{i \in A_{st}} h_{it}$$

- We propose the following specification, *i.e.*, **Model A**,

$$\ell_{it} = \alpha + \beta_1 h_{it} + \beta_2 \bar{h}_{st} + X_{it}\delta + \lambda_s + v_{it},$$

- with employee fixed effects,

$$v_{it} = u_i + \epsilon_{it}.$$

Econometrics 2 : Group Synergies ; District Level

- Interpretation : if we aggregate over all i in subset A_{st}, we find,

$$\bar{\ell}_{st} = \alpha + (\beta_1 + \beta_2)\bar{h}_{st} + \bar{X}_{st}\delta + \bar{v}_{st},$$

- If we measure economies of density at the level of the employee, we have $\gamma = 1 - \beta_1$.
- To measure economies of density at the district level, we have

$$\gamma = 1 - \beta_1 - \beta_2.$$

- Table 2 below shows that $\gamma = 1 - \beta_1 - \beta_2 \simeq 0.42$.

Table 1 : OLS and Fixed Effects Estimation. Model A

	(1) OLS	(2) OLS	(3) OLS	(4) FE
h_{it}	1.447***	1.404***	**1.451*** **	**1.109*** **
	(0.020)	(0.021)	(0.020)	(0.018)
$\bar{h}_{st}$	-0.795***	-0.767***	**-0.867*** **	**-0.522*** **
	(0.032)	(0.042)	(0.044)	(0.028)
Constant	1.828***	2.457**	1.261	2.128***
	(0.168)	(0.696)	(0.252)	(0.137)
Controls	NO	YES	NO	NO
District Indicators	NO	YES	YES	NO
$\beta_1 + \beta_2$	0.652	0.637	**0.584**	**0.587**
Observations	56,878	56,830	56,878	56,878
Groups	.	.	.	3,688
R^2	0.268	0.553	0.488	.

Dependent variable : Log-Kilometers ℓ_{it}. Column (4) gives the within-group, fixed effects estimation, with employees i as groups.

Econometrics 3 : IVs and Arellano-Bond GMM estimates

- We push the analysis further : estimate the model in first differences and use lagged hours as IVs. To take care of possible correlations of first-differenced shocks with hours.
- We estimate a model with two lags of the dependent variable,

$$\ell_{it} = \rho_1 \ell_{i,t-1} + \rho_2 \ell_{i,t-2} + \beta_1 h_{it} + \delta_1 h_{i,t-1} + \beta_2 \bar{h}_{st} + \delta_2 \bar{h}_{s,t-1} + u_i + \epsilon_{it},$$

- We measure economies of density at the long-run stationary equilibrium. We have,

$$\gamma = 1 - \frac{\beta_1 + \beta_2 + \delta_1 + \delta_2}{1 - \rho_1 - \rho_2}.$$

- Table 3 below shows that $\gamma \simeq 0.7$.
- We find $\rho_1 \simeq .1$ and $\rho_2 \simeq .02$. Autocorrelation is not strong. $h_{i,t-1}$ is not significant.

Table 2 : Arellano-Bond GMM estimates

	(1)	(2)	(3)	(4)
ℓ_{it-1}	0.112^{***}	0.111^{***}	0.085^{***}	0.090^{***}
	(0.021)	(0.019)	(0.020)	(0.020)
ℓ_{it-2}	0.020^{*}	0.023^{**}	0.010	0.011
	(0.010)	(0.009)	(0.009)	(0.009)
h_{it}	1.159^{***}	1.268^{***}	1.216^{***}	1.140^{***}
	(0.054)	(0.068)	(0.062)	(0.066)
$\bar{h}_{st}$	-1.089^{***}	-1.261^{***}	-1.194^{***}	-1.1097^{***}
	(0.070)	(0.089)	(0.085)	(0.087)
$\bar{h}_{st-1}$	0.249^{**}	0.202^{*}	0.209^{**}	0.258^{**}
	(0.085)	(0.094)	(0.087)	(0.088)
γ	**0.662**	**0.706**	**0.737**	**0.716**
lags of ℓ_{it} used as IV	all	3	10	10
lags of h_{it} and $\bar{h}_{st}$ used as IV	3	3	3	10
Autocorr. test order 1 (p-value)	0.000	0.000	0.000	0.000
Autocorr. test order 2 (p-value)	0.693	0.340	0.635	0.825
GMM Steps	1	1	2	2

Motivation for the Use of a Finite Mixture Model

- We suspect that the average transportation-cost functions L/H may in fact be U-shaped, as textbook AC curves.
- These curves can be fitted by a quadratic model in logs.
- If we try to estimate a quadratic extension of our model, we find that quadratic terms are non-significant and useless.
- If we try to model unobserved heterogeneity by assuming the existence of a finite number of latent types, then, we find that the quadratic model is relevant.

Latent Types Reveal Unobserved Heterogeneity and Treat Endogeneity Problems

- The estimation of a model with latent groups of employees reveals the unobserved heterogeneity of the workers.
- Some unobservable employee types, but not all, are responsible for the economies of density observed in the aggregate.
- Adding observable controls to the model is not enough. For instance, latent types are still relevant, both in rural and urban districts. Types are not simply capturing the impact of rural vs. urban location.
- In other words : A classification of employees based on observable characteristics would not uncover the structure of density economies.
- Finally, the use of latent types is a way of treating potential endogeneity problems : *we assume that error terms are independent of hours conditional on the latent type*.

Econometrics 4. A Quadratic Model with Unobserved Types (Finite Mixture)

- We assume that each individual i at date t may belong to one of K unobservable groups indexed by $k = 1, \ldots, K$.
- For an (i, k) in group k, we assume that the following model, called **MODEL B**, describes traveled distance,

$$\ell_{it} = \alpha_k + \beta_k \bar{h}_{st} + \nu_k \bar{h}_{st}^2 + \delta_k h_{it} + \rho_k (1/n_{st}) + \epsilon_{itk},$$

  and $\epsilon_{itk} \sim \mathcal{N}(0, \sigma_k^2)$.
- We want to estimate parameters $(\alpha_k, \beta_k, \nu_k, \delta_k, \rho_k, \sigma_k)$ for each k, and the prior probability of type k, denoted p_k.
- So ℓ_{it} is distributed like a **mixture of normal distributions**.
- n_{st} is the number of employees in district s at time t.
- Key assumption : $\mathbb{E}(\epsilon \mid h, n, k) = 0$.

Econometrics 4. Interpretation of Model B

- Define the log-ratio of kilometers to hours of service of district s at time t as follows : $\bar{\kappa}_{st} = \bar{\ell}_{st} - \bar{h}_{st}$.
- We aggregate the model over all $i \in A_{st}$ (for a fixed type k) and we find,

$$\bar{\kappa}_{st} = \alpha + (\beta + \delta - 1)\bar{h}_{st} + \nu \bar{h}_{st}^2 + \rho(1/n_{st}) + \bar{\epsilon}_{st},$$

  dropping index k to lighten notation.
- For each type $k = 1, \ldots, K$, we compute the expectation of the average ratio $L/H = \exp(\bar{\kappa})$, using the estimated values of the parameters.

Computation of the L/H-curves (i.e., U-shaped curves)

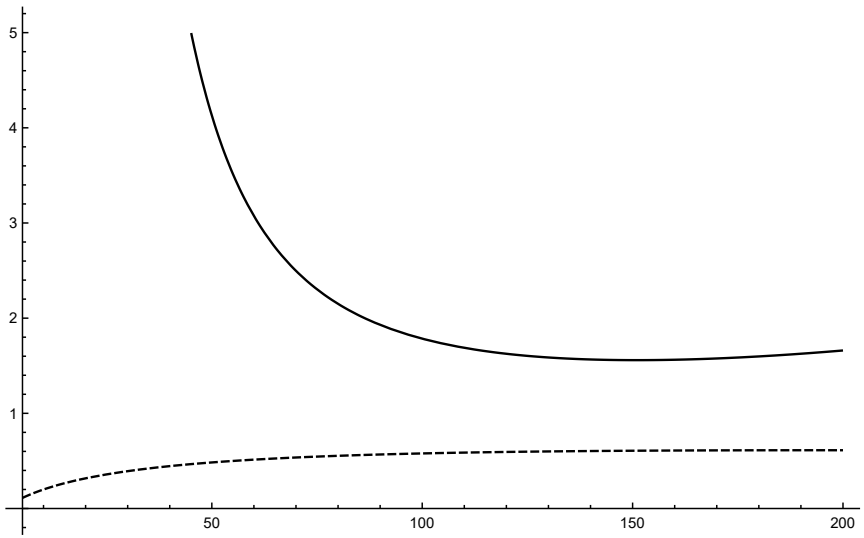
- If we assume that the ϵ_{itk} are normal, i.i.d., with mean zero and variance σ_k^2, we have,

$$\mathbb{E}\left(e^{\bar{\kappa}} \,\middle|\, \bar{h}, k, n\right) = \exp\left\{\alpha_k + (\beta_k + \delta_k - 1)\bar{h} + \nu_k \bar{h}^2 + \left(\rho_k + \frac{\sigma_k^2}{2}\right)\frac{1}{n}\right\},$$

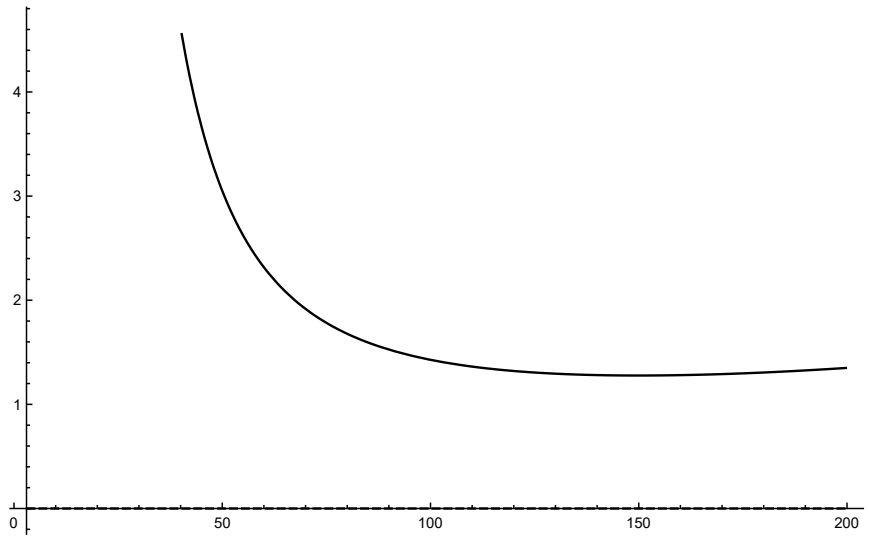
- We see that $\bar{\kappa}$ is given by a quadratic curve in $\bar{h}$.
- To obtain the average L/H curve, we just use the estimated values of p_k to compute,

$$\mathbb{E}\left(\exp(\bar{\kappa})|\bar{h}, n\right) = \sum_k p_k \mathbb{E}\left(\exp(\bar{\kappa})|\bar{h}, n, k\right).$$

Figure 3. L/H Curves Conditional on Type k and Average L/H Curve, for $K = 2$

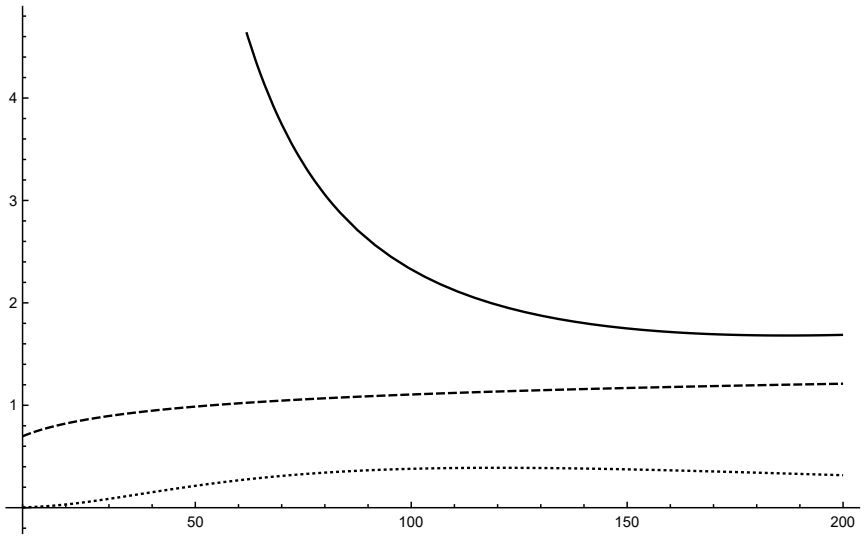


a. Type 1 (solid line); Type 2 (dotted line)

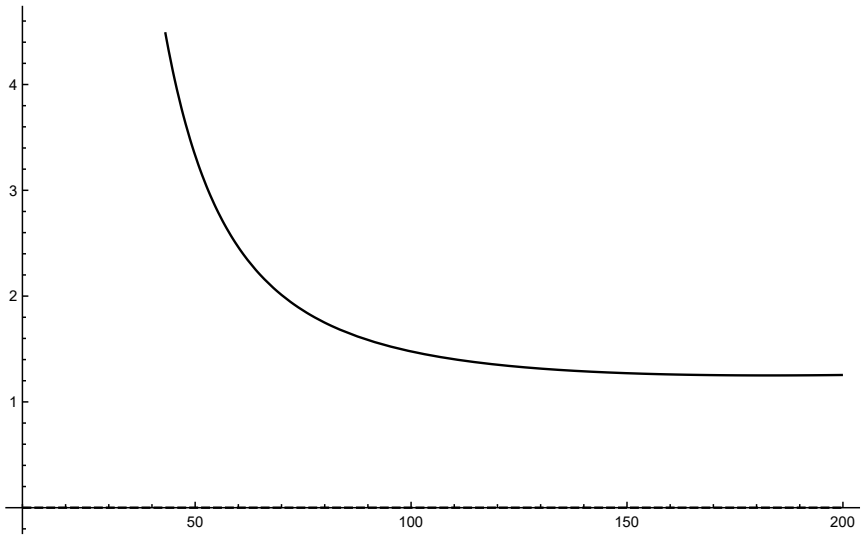


b. Average L/H curve

Figure 4. L/H Curves Conditional on Type k, for $K = 3$

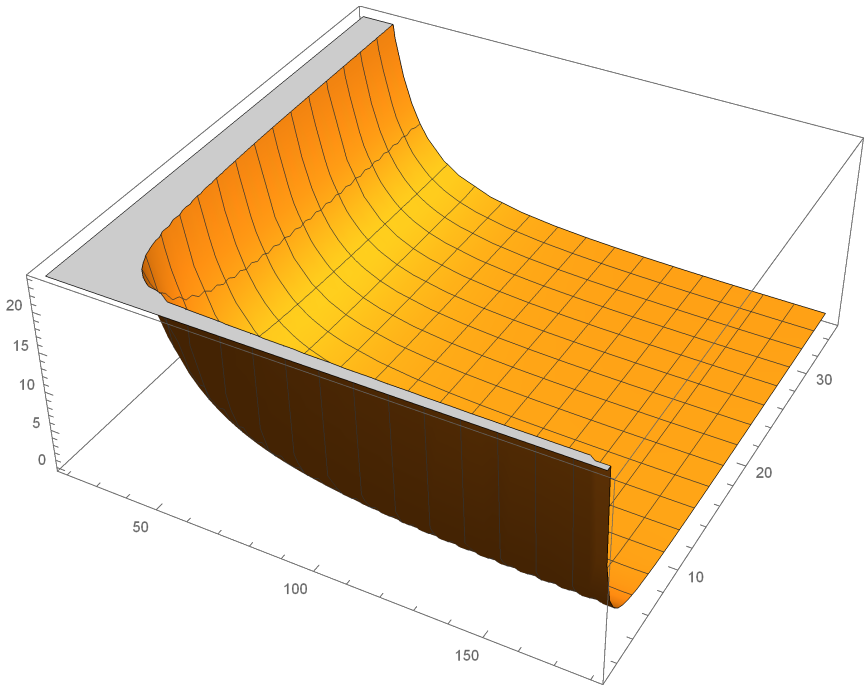


a. Type 1 (solid line), Type 2 (dashed line), Type 3 (dotted line)

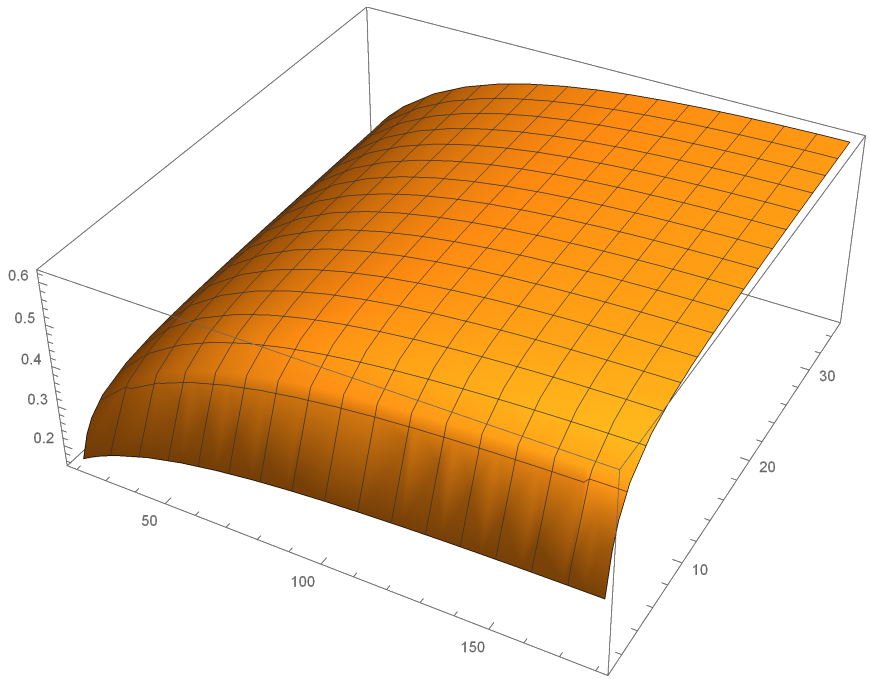


b. Average L/H curve

Figure 5. 3D Plots of L/H surfaces conditional on type k, for $K = 2$

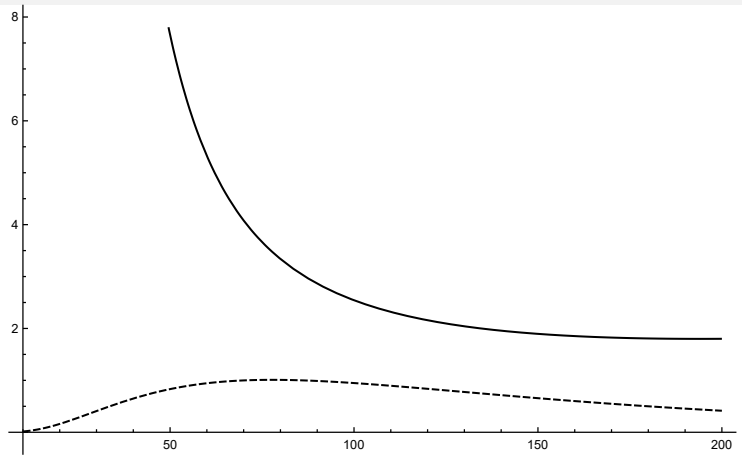


a. Type 1

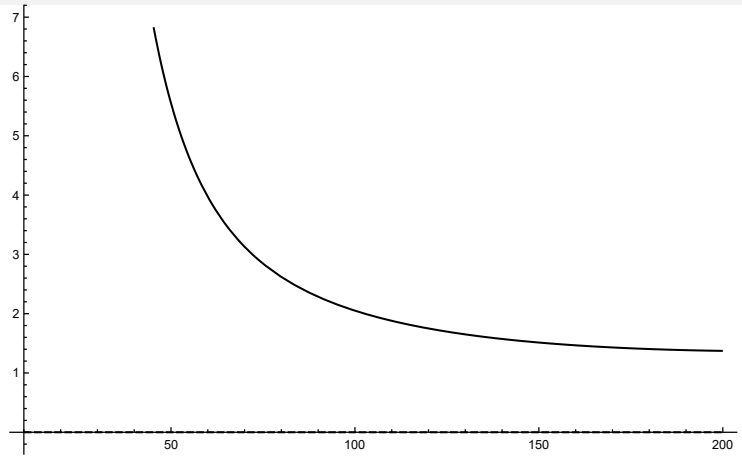


b. Type 2

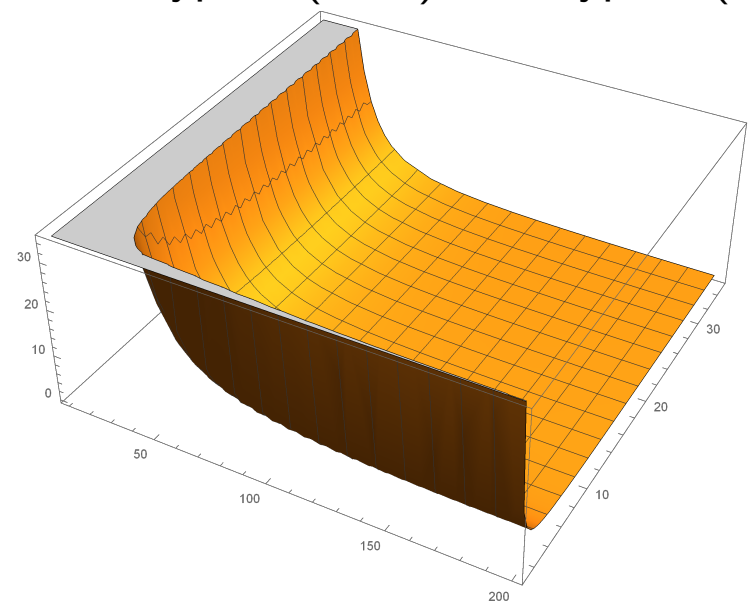
Figure 6. L/H Surfaces in Rural Districts, $K = 2$



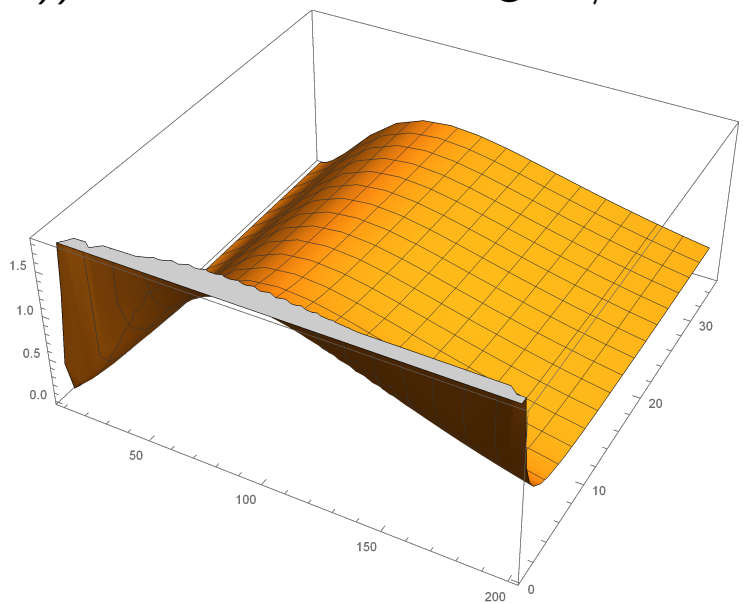
a. Rural Type 1 (solid) and Type 2 (dashed))



b. Rural Average L/H curve

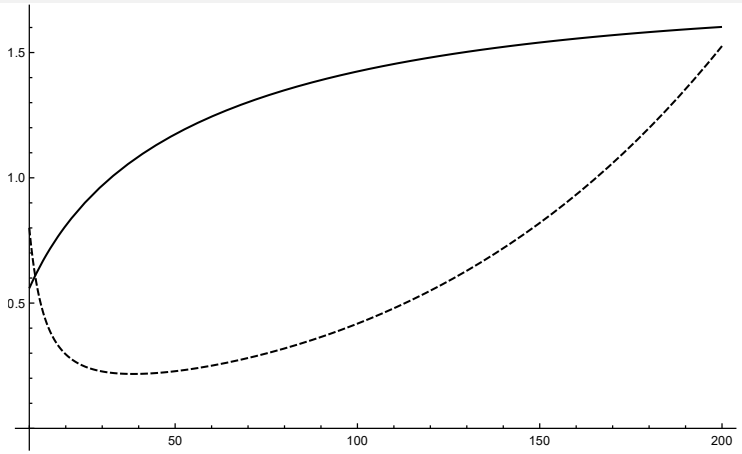


c. Rural Type 1

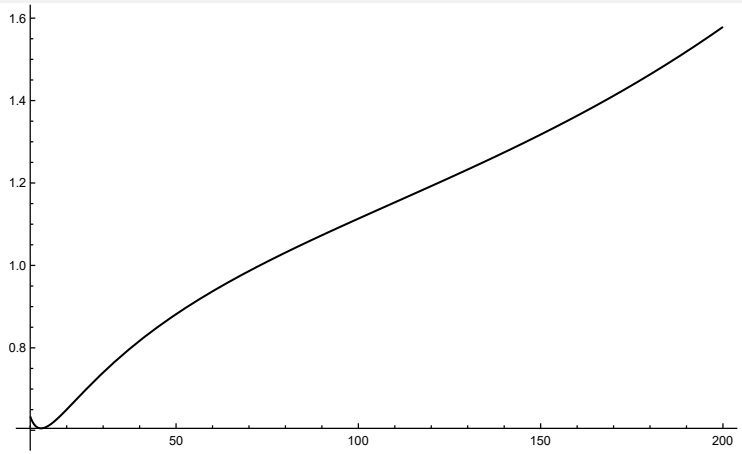


d. Rural Type 2

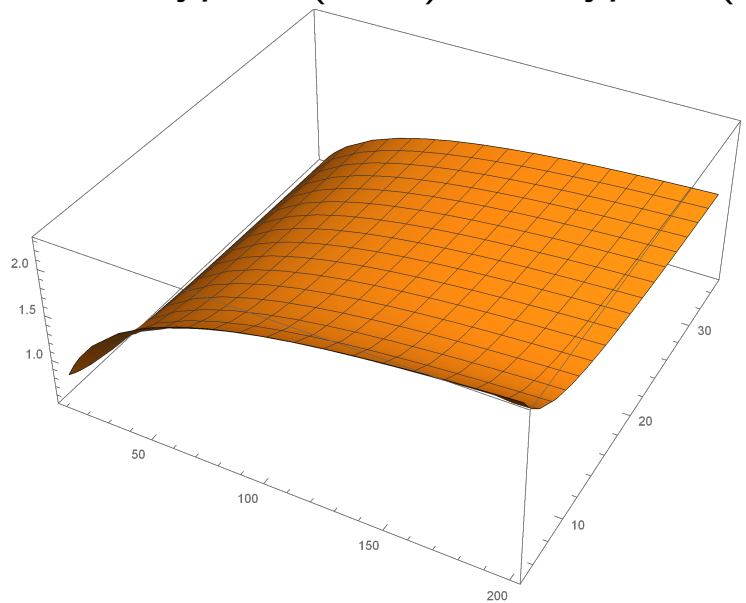
Figure 7. L/H Surfaces in Urban Districts, $K = 2$



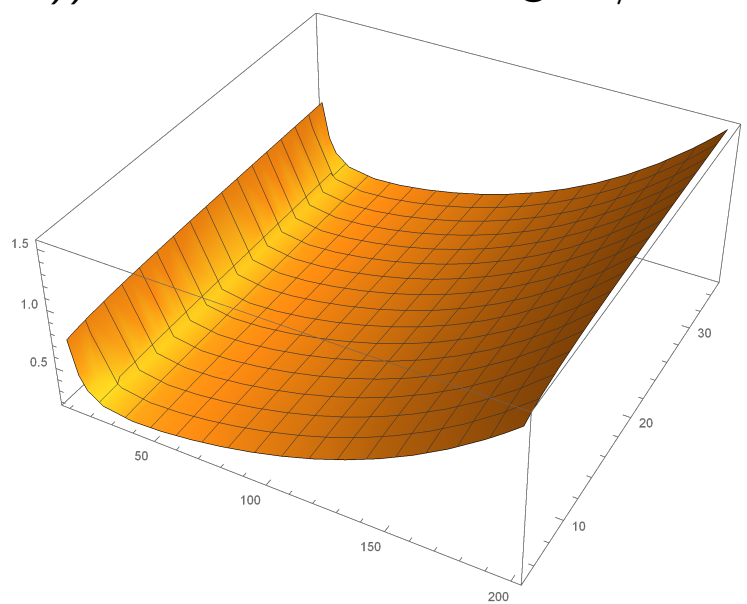
a. Urban Type 1 (solid) and Type 2 (dashed))



b. Urban Average L/H curve

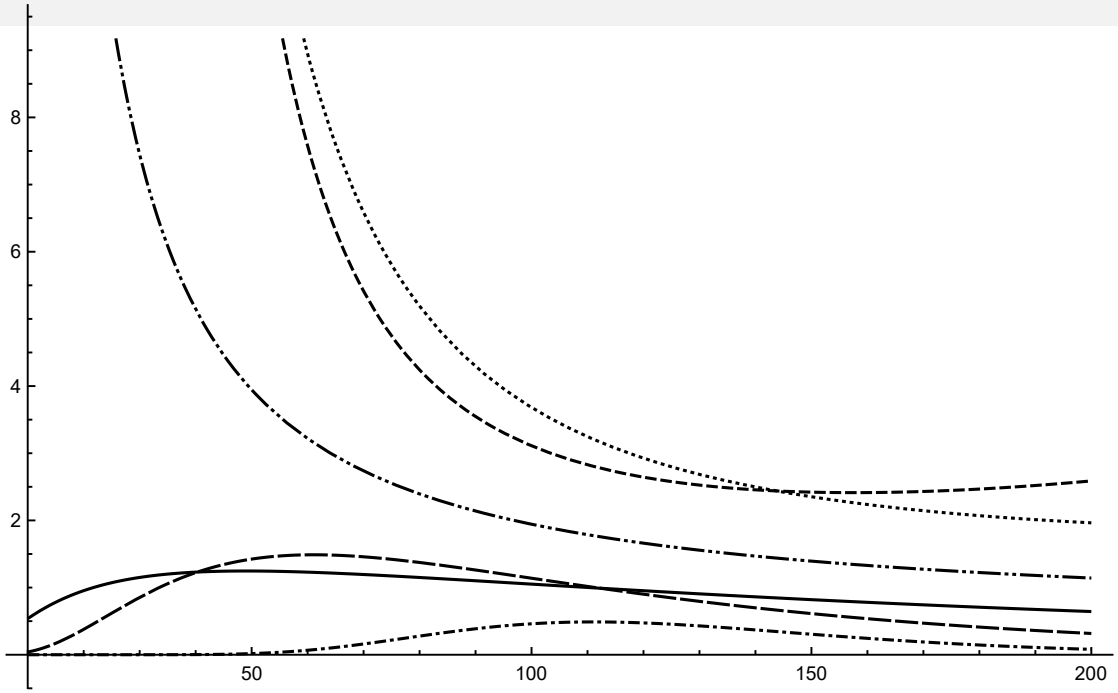


c. Urban Type 1

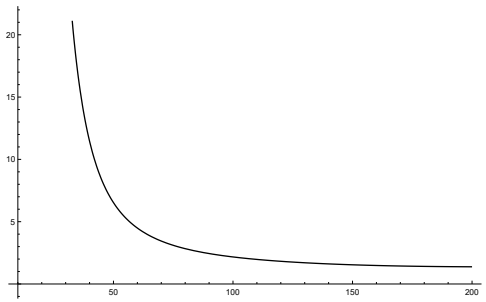


d. Urban Type 2

Figure 8. L/H curves in Rural Districts and $K = 6$



a. Type 1 (solid), Type 2 (dashed), Type 3 (dotted), Type 4 (-·-·), Type 5 (long -), Type 6 (-··-··).



b. Average L/H curve over 6 types

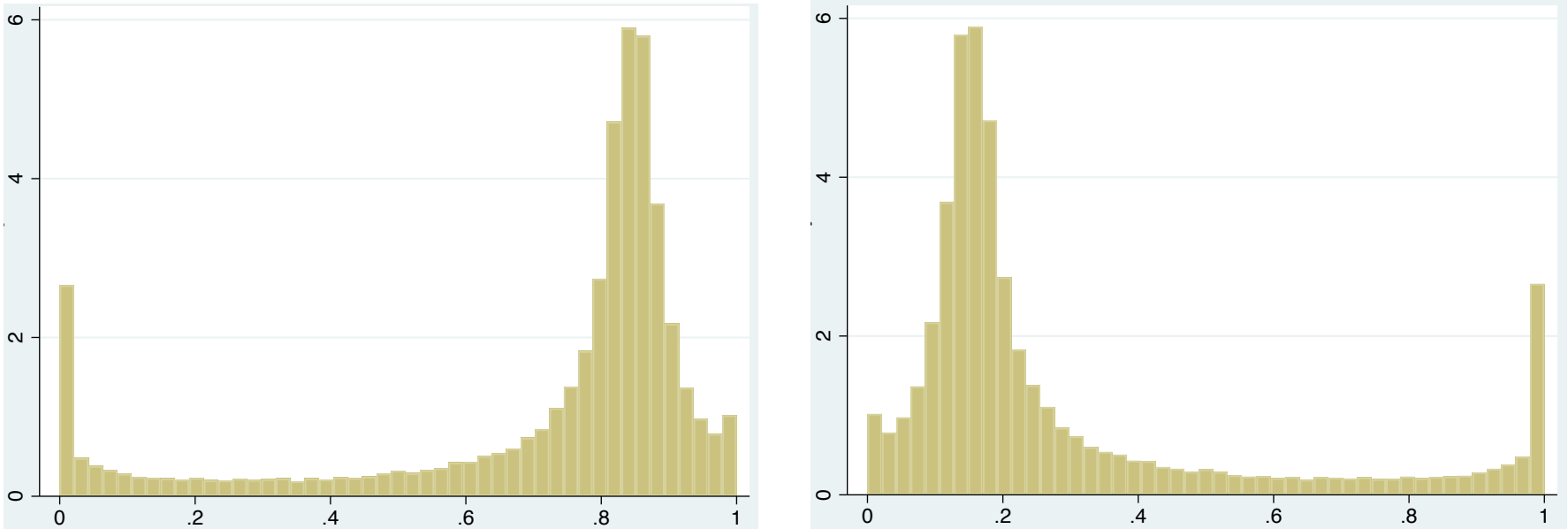
Choice of the Number of Types
Quality of Classification (Visual Inspection)

- The log-likelihood always increases with K (but the curve is concave as a function of K).
- To assess the quality of classification, we focus on the posterior probabilities of types, computed with the help of the likelihood function and ML estimates. We define,

$$p_{itk} = \Pr[it \in k \,|\, \ell_{it}, h_{it}, X_{it}].$$

- Visual inspection of the distribution (histogram) of $\hat{p}_{itk}$ for each type k gives a good idea of the quality of classification...

Figure 9. Posterior Probabilities of Types : Density of Estimated Posterior Probabilities of Types $\hat{p}_{it1}$ for $K = 2$

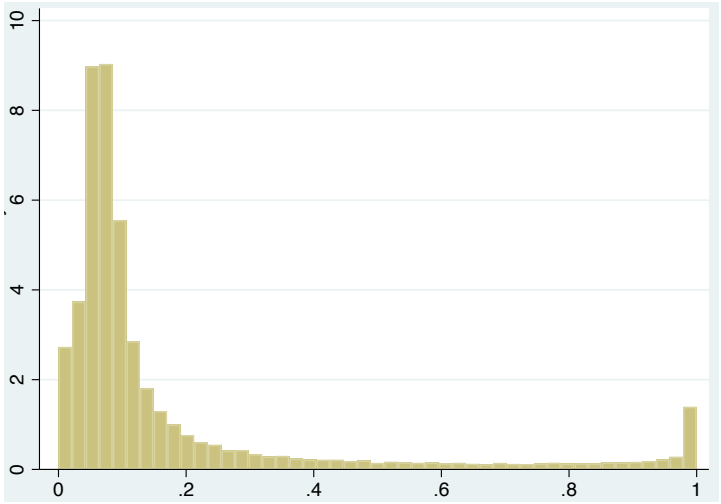


a. Distribution of p_{it1} (Type 1)

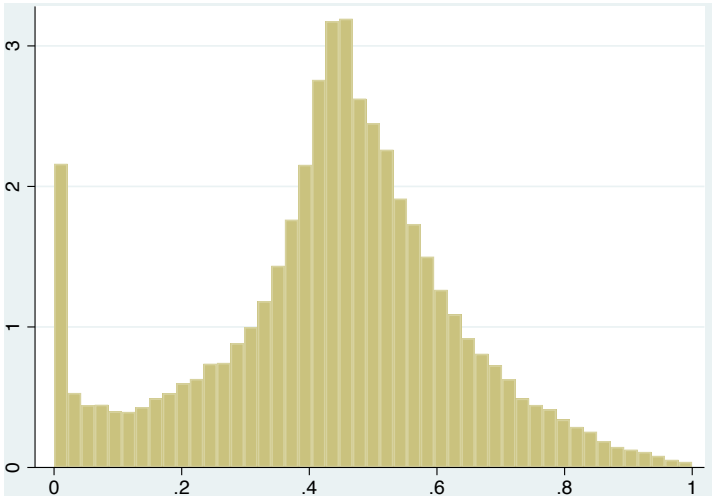
b. Distribution of p_{it2} (Type 2).

The classification of types is reasonably good for $K = 2$ in the Rural and Urban districts sub-samples. This is also true in the full sample including all districts.

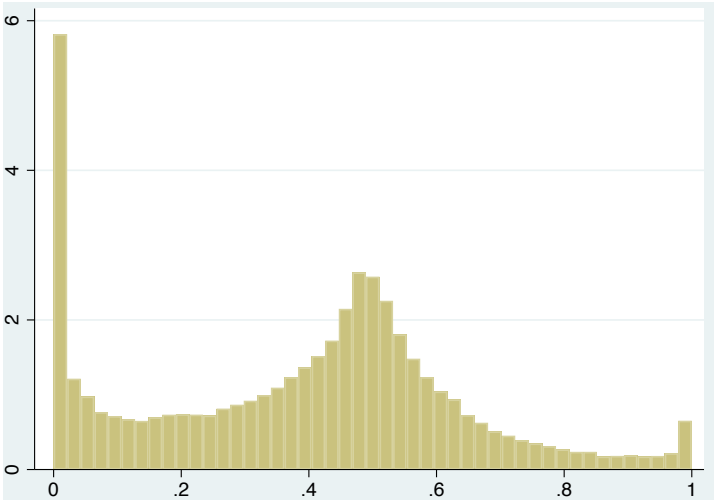
Fig. 10. Estimated Posterior Probabilities of Types, $K = 3$



a. Distribution of p_{it1} (Type 1)



b. Distribution of p_{it2} (Type 2).



c. Distribution of p_{it3} (Type 3).

Conclusion

- We studied observations of a network of Home Care Services (in panel form).
- Standard econometric techniques have revealed the existence of economies of density : transportation cost per hour of service decreases when hours of service increase at the district level.
- Economies of density are the result of group synergies at the level of local branches (*i.e.*, districts).
- The finite mixture approach to estimation shows that the observed U-shaped average cost curves are in fact an average of type-dependent curves.
- We find that the best modelling choice is to keep only two types.
- One of the two types only is responsible for the economies of density observed in the aggregate.

Fig. A1. Km/Hours of Service Ratio at the County (*i.e.*, *département* Level

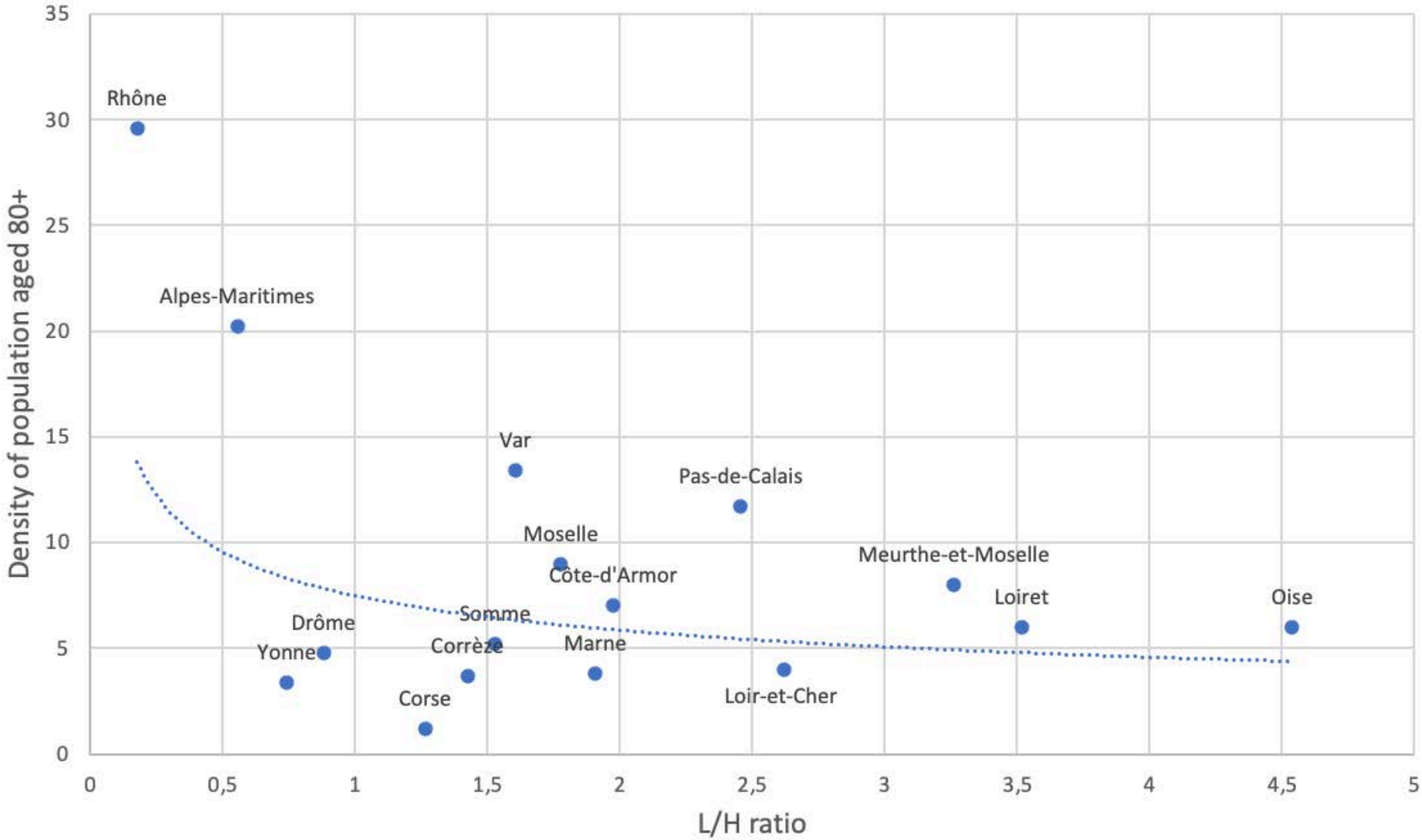


Table A1 : OLS. Pooled Data. Log-Travelled Kilometers ℓ_{it}

	(1)	(2)	(3)	(4*)
h_{it}	**1.013*****	**1.049*****	.	.
	(0.025)	(0.025)		
$h_{it} *$ Full Time	.	.	**1.127*****	**1.255*****
			(0.037)	(0.045)
$h_{it} *$ 80%Time	.	.	**1.046*****	**1.113*****
			(0.022)	(0.027)
$h_{it} *$ Part Time	.	.	**0.972*****	**0.973*****
			(0.040)	(0.052)
Full Time	**0.663*****	**0.642*****	-0.070	-0.692*
	(0.023)	(0.024)	(0.250)	(0.307)
80%Time	**0.470*****	**0.392*****	0.146	-0.222
	(0.018)	(0.018)	(0.197)	(0.262)
Constant (ref. Part Time)	-0.259*	-0.259	-0.093***	0.601
	(0.104)	(0.572)	(0.168)	(1.784)
District dummies & controls	NO	YES	NO	YES
Observations	56,878	56,830	56,878	56,830
*Clusters	no	no	no	3686
R^2	0.28	0.57	0.28	0.57

A2 Econometrics : Panel Analysis at the District Level

- If we aggregate (take averages of) all kilometers and hours at the district level (*i.e.*, we average variables over all employees i belonging to district s at time t), we obtain the model,

$$\bar{\ell}_{st} = \alpha + \beta_3 \bar{h}_{st} + \bar{X}_{st}\delta + \bar{u}_{st},$$

- where $\bar{\ell}_{st}$ denotes the average traveled kms of employees, per month, in district s, month t.
- and where $\bar{h}_{st}$ is the average hours of service of employees, per month, in district s, month t...
- *Interpretation* : $\beta_3 \simeq \beta_1 + \beta_2$, and β_3 represents *team synergies* at the district level.

Table A2 : District-Level Panel-Data Estimates

	(1) OLS	(2) OLS	(3) Fixed Effects	(4) First Differences
$\overline{h}_s$	0.774***	.	.	.
	(0.110)	(.)	(.)	(.)
$\overline{h}_{st}$	.	**0.498****	**0.515****	.
	(.)	(0.046)	(0.051)	(.)
$\overline{h}_{st} - \overline{h}_s$	**0.502****	.	.	.
	(0.097)	(.)	(.)	(.)
$\Delta\overline{h}_{st}$	.	.	.	0.429***
	(.)	(.)	(.)	(0.057)
Constant	1.265**	1.641***	2.465***	-0.000
	(0.513)	(0.217)	(0.237)	(0.005)
District Dummies	NO	YES	NO	NO
Observations	3,117	3,117	3,117	3,065
Groups	.	.	98	.
R^2	0.027	0.891	.	0.072

Note : Regressions (1), (2) and (4) are weighted, with weights n_{st} equal to the number of employees in district s in period t. The within estimator of column (3) is also weighted, by the average number $\bar{n}_s$ of employees in district s.

A3. The Beardwood-Halton-Hammersley Theorem (1959)

- The Beardwood-Halton-Hammersley Theorem is a famous result in Applied Probability and Combinatorial Optimization.
- Suppose there are ν points, $(x_1, \ldots, x_\nu)$ drawn at random and i.i.d. on $[0,1]^2$. Let $L(x_1, \ldots, x_\nu)$ denote the minimal length in the set of all tours joining the ν points.
- *BHH Theorem*. If (x_i), $i = 1, \ldots, \nu$ are i.i.d. and distributed on $X \subset [0,1]^2$ with a nonzero area, then there exists a constant ρ s.t., with probability 1,

$$\frac{L(x_1, \ldots, x_\nu)}{\sqrt{\nu}} \to \rho \quad \text{as the number of visits} \to \infty.$$

- This suggests that if the carers' tours are organized efficiently, then,

$$\frac{L}{H} \simeq \frac{A(X)}{\sqrt{H}} \quad \text{or, taking logs,} \quad \ell = \alpha(X) + \frac{h}{2} + \epsilon.$$

Table A3 : "Test" of the BHH Theorem. Total hours and Km

	(1) OLS	(2) OLS	(3) FE
Dependent Variable :	$\ln(\sum_{t\in T_i} L_{it})$	$\ln(\sum_{i\in A_{st}} L_{it})$	$\ln(\sum_{i\in A_{st}} L_{it})$
$\ln(\sum_{t\in T_i} H_{it})$	0.987***	.	.
	(0.016)	(.)	(.)
$\ln(\sum_{i\in A_{st}} H_{it})$	.	**0.438*** **	**0.561*** **
	(.)	(0.035)	(0.090)
Constant	-0.504*	4.130***	3.897***
	(0.195)	(0.274)	(0.699)
District Dummies	YES	YES	NO
Observations	3,687	3,117	3,117
Groups	.	.	98
R^2	0.831	0.950	.

Note : Regressions (2) and (3) are weighted. Weights are n_{st} for (2) and $\bar{n}_s$ for (3).

Econometrics A4. Log-Likelihood

- First we write the contribution to likelihood of (i, t) with avatar k,

$$\Lambda_{itk} = \left(\frac{1}{\sigma_k}\right) f\left(\frac{\epsilon_{itk}}{\sigma_k}\right),$$

  where f is the standard normal density.
- (i, t)'s contribution to likelihood is then

$$\Lambda_{it} = \sum_{k=1}^{K} p_k \Lambda_{itk}.$$

- The log-likelihood can be written,

$$\ln \Lambda = \sum_{i=1}^{N} \sum_{t \in T_i} \ln \left(\sum_{k=1}^{K} p_k \Lambda_{itk} \right),$$

  where T_i is the set of dates t such that i is observed.

Econometrics A4. Estimation by EM and ML

- We use a sequential EM algorithm to obtain preliminary estimates of all parameters. The model is then estimated by straightforward Maximum Likelihood.
- The EM algorithm is of the type discussed in the work of Jean-Marc Robin and others. (See *e.g.*, Arcidiacono and Jones (2003), Bonhomme and Robin (2009), Gary-Bobo, Goussé and Robin (2016).)
- A side-product of the algorithm is the *classification* of each (i, t), given by p_{itk}. We apply Bayes's rule to obtain the *posterior* probabilities :

$$p_{itk} = \frac{p_k \Lambda_{itk}}{\sum_{j=1}^{K} p_j \Lambda_{itj}} = \Pr[it \in k | h_{it}, \ell_{it}, X_{it}].$$

- We find that it is not useful to estimate more than three types : we have $K = 2$ or $K = 3$.

Econometrics A4. ML estimates of the quadratic model

- To estimate prior probabilities, we use a classic parametrization. Define $(r_1, \ldots, r_K)$ such that

$$p_k = \frac{e^{r_k}}{\sum_{j=1}^{K} e^{r_j}} \qquad \text{with} \qquad r_1 = 0.$$

- We present first ML estimates for one two or three types. A huge increase in the likelihood is achieved when we move from $K = 1$ to $K = 2$.
- We also estimated the model with two distinct sub-samples : *urban* and *rural* districts (results presented below).

Table A4. ML Estimation of Model B with $K = 1, 2, 3$

(K, k)	(1,1)	(2,1)	(2,2)	(3,1)	(3,2)	(3,3)
$\bar{h}_{st}$	-0.168	-8.125^{***}	0.071	-8.666^{***}	-0.375	5.908^{**}
	(0.600)	(0.812)	(0.915)	(1.088)	(0.670)	(2.081)
$\bar{h}^2_{st}$	-0.067	0.798^{***}	-0.127	0.830^{***}	-0.023	-0.786^{**}
	(0.066)	(0.088)	(0.102)	(0.119)	(0.074)	(0.229)
h_{it}	1.460^{***}	1.118^{***}	2.268^{***}	0.983^{***}	1.734^{***}	2.616^{***}
	(0.010)	(0.010)	(0.033)	(0.012)	(0.025)	(0.052)
$1/n_{st}$	-0.447^{***}	0.522^{***}	-1.468^{***}	3.220^{***}	-0.173	-1.857^{***}
	(0.096)	(0.105)	(0.233)	(0.189)	(0.106)	(0.329)
Constant	0.323	20.495^{***}	-3.987	23.076^{***}	-1.068	-18.901^{***}
	(1.376)	(1.876)	(2.062)	(2.497)	(1.513)	(4.774)
σ_k	1.057^{***}	0.688^{***}	1.211^{***}	0.590^{***}	0.679^{***}	1.262^{***}
	(0.003)	(0.004)	(0.009)	(0.006)	(0.010)	(0.015)
r_k	.	0	-0.866^{***}	0	0.069	-0.886^{***}
	(.)	(0)	(0.034)	(0)	(0.045)	(0.061)
p_k	1	0.704	0.296	0.403	0.431	0.166
Log-Lik	$-84,254$	$-79,507$		$-78,522$		

Choice of the Number of Types
Quality of Classification

- The choice of the appropriate number of types K is a delicate question.
- The log-likelihood always increases with K (but the curve is concave as a function of K).
- We can use the usual information criteria : the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).
- AIC tends to overestimate K (overfitting) and BIC tends to underestimate K.
- AIC and BIC do not penalize a model for a bad *quality of classification* when K is large ...
- so, we will also consider some *Entropy Criteria* (below).
- Visual inspection of the distribution (histogram) of $\hat{p}_{itk}$ for each type k gives a good idea of the quality of classification...

Entropy Criteria

- The entropy of the model's classification is defined as follows,

$$\mathcal{E}(K) = -\sum_{i}\sum_{t\in T_i}\sum_{k=1}^{K}\hat{p}_{itk}\ln(\hat{p}_{itk}).$$

- We can divide entropy by its maximum value $N\ln(K)$, yielding

$$0 \leq \mathfrak{E}(K) = \frac{\mathcal{E}(K)}{N\ln(K)} \leq 1.$$

- Celeux and others (*e.g.*, Celeux and Soromenho (1996)) have suggested the Normalized Entropy Criterion, defined for $K > 1$,

$$NEC(K) = \frac{\mathcal{E}(K)}{\ln\Lambda(K) - \ln\Lambda(1)}.$$

Table A5. Information and Entropy Criteria : Choice of K

K	1	2	3	5	6	7	8
$L(K)$	-82,468	-77,013	-75,819	-74,718	-74,497	-74,321	-74,143
BIC	165,112	154,388	152,186	150,357	150,100	149,935	**149,765**
AIC	164,698	154,093	151,738	149,605	149,196	148,880	148,557
$\mathcal{E}(K)$	0	**24,132**	43,992	64,836	73,234	76,956	82,018
NEC	.	**4.424**	6.616	8.366	9.187	9.446	9.852
$\mathfrak{E}(K)$	0	**0.609**	0.701	0.705	0.715	0.692	0.690
AHHI	1	**0.733**	0.529	0.397	0.345	0.330	0.306
$\mathfrak{H}(K)$	.	**0.466**	0.293	0.246	0.214	0.218	0.206

Note : Model B has been estimated by the EM algorithm repeatedly with values of K ranging from $K = 1$ to $K = 9$. Model B has been estimated, adding interactions of the K types with the indicators of three subsamples : the Urban, Peri-Urban and Rural districts (that partition the dataset). Parameters therefore vary not only with type, but also with the three types of district.
$L(K)$ is the estimated Log-Likelihood with K types. BIC is the Bayesian Information Criterion. AIC is Akaike's Information Criterion. $\mathcal{E}(K)$ is entropy as defined above.
AHHI is the Average Hirschman-Herfindahl Index defined in the text.
$\mathfrak{H}(K)$ is the normalized Herfindahl index defined as follows.

$$\text{AHHI}(K) = \frac{1}{N} \sum_{i=1}^{n} \sum_{t \in T_i} \sum_{k=1}^{K} \hat{p}_{itk}^2$$

In the case of Model B ; BIC seems to reach a minimum for $K = 8$; AIC never reaches a minimum between $K = 1$ and $K = 9$; NEC is minimal for $K = 2$. Most of the gains in terms of $L(K)$ are achieved with $K = 2$ or $K = 3$.