

Does neighborhood matter? The impact of proximity to (dis)amenities on home price

Thierry Kamionka (CNRS and CREST)



Introduction

- The increase in interest rates during 2022, 2023 and the first semester 2024 has resulted in a slowdown in the real estate transactions and a reduction of activity in the construction section in France.
- Price of housing can depend on physical characteristics and neighborhood characteristics.
- Amenities in the literature, some examples : transportation, school quality, air pollution, sea level rise, crime, EPC.
- Amenities and disamenities can be correlated (train stations, green spaces, energy efficiency distribution).

Introduction

- Among these (dis)amenities, noise level (road, train, airplane) or delinquency may be important determinants of transaction prices.
- We model prices of apartments in Paris by enriching the data with neighborhood characteristics.
- We use hedonic models as well as mixture of hedonic models. We take into account the existence of clusters at the district level.
- We consider potential links between the modeling we use and spatial econometrics.
- We examine a variant of modeling such that the distribution of unobserved heterogeneity may be correlated to other explanatory variable(s) .

What are our objectives?

- Enrich a set of geocoded administrative data on real estate transactions with information on (dis)amenities.
- Including noise levels, train or subway stations, educational establishments, burglary rates (geocoded), Energy efficiency label.
- Model unobserved heterogeneity in such a way that distribution is neighborhood specific.
- Use an estimation method adapted to this type of mixture models.
- Estimate the models, use the panel data dimension, nonlinearities in effects.
- Study some link that may exist between panel data econometrics and spatial econometrics.

What have we found?

- Price of apartments have decreased in 2022, 2023 and 2024.
- Characteristics of apartment matters: surface (non linear effect), number of rooms, presence of a dependence, energy efficiency label.
- Distance of the center of Paris has a large and negative impact (historical development of Paris).
- Distance to the nearest station (train, RER, subway): monotone impact (price increases with distance, 303 subway stations).
- Negative effect of device proximity of the ring road ('périphérique'). Air pollution (particulate matter, PM2,5), traffic congestion.
- Passenger traffic at the nearest station has large and negative impact of the transaction price.

What have we found?

- Two have two or three connections at the nearest station has a negative impact (more thefts?, Montparnasse, Gare du Nord, Gare de Lyon).
- Noise pollution has a significant impact (negative from 45 to 60 dB). Particular case prestigious avenues, the banks of the Seine.
- Delinquency (burglary rates) has a significant and large impact (informational of the 'quality' of apartments).
- Distance to the nearest high school has a negative and large impact.
- Distance to the nearest 'best' high schools has a negative impact.
- Energy efficiency label matter. Importance of quality. However, effects are less marked than they probably are in some regions for very low-energy-performance housing.

- The logarithm of the transaction price for an appartement y_i is given by

$$y_i = x_i' \beta + \alpha_{q(i)} + u_i,$$

where β is a vectors of parameters, u_i is an error term, $\alpha_{q(i)}$ is an individual effect for district $q(i)$ where transaction i took place.

- $\alpha_{q(i)}$ is a fixed effect or a random effect. For the specifications such that $\alpha_{q(i)}$ is a random effect, we consider two variants: continuous unobserved heterogeneity and discretized unobserved heterogeneity.

The logarithm of the transaction price of an apartment for transaction i is given by the following mixture hedonic price models

$$y_i = x_i' \beta + \gamma_{k(i)} + u_i$$

where x_i is a vector of observable characteristics for the transaction i ,

γ_k is an unobserved characteristics ($k \in \{1, \dots, K\}$) and the error term $u_i \sim N(0; \sigma^2)$.

For a given neighborhood q , the proportion of transactions that belong to the category k is denoted $\delta_{k,q}$, where $\sum_{k=1}^K \delta_{k,q} = 1$ and $\delta_{k,q} \in [0; 1]$.

The conditional density of the logarithm of the price given observed characteristics x_i and neighborhood $q(i)$ is

$$f(y_i | x_i, q(i); \theta) = \sum_{k=1}^K \delta_{k,q(i)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - x_i'\beta - \gamma_k)^2}{2\sigma^2}\right) \quad (1)$$

$\theta = (\beta, \gamma, \delta, \sigma^2)$ is the vector of parameters.

→ as a function of θ it is a contribution to the likelihood function.

Some characteristics z of the apartment is unobserved and the distribution of this component within a given neighborhood q is given by the vector of probabilities $\delta_q = (\delta_{1,q}, \dots, \delta_{K,q})'$.

In such a context, we can consider that we are in presence of an incomplete sampling scheme.

Let $y_i^* = (y_i, z)$ denote the vector of the logarithm of the transaction price and the unobserved component. The conditional density of y_i^* is

$$f(y_i^* \mid x_i, q(i); \theta) = \delta_{z,q(i)} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(y_i - x_i' \beta - \gamma_z)^2}{2\sigma^2}\right) \quad (2)$$

This model can be estimated using an **EM algorithm** (see Dempster, Lair and Rubin, 1977).

Let us consider the following function of the vector of parameters θ

$$Q(\theta \mid \dot{\theta}) = \frac{1}{n} \sum_{i=1}^n E_{\dot{\theta}}[\log(f(y_i^* \mid x_i, q(i); \theta)) \mid y_i, x_i, q(i)]$$

where the expectation is defined with respect to the conditional density of y_i^* given $y_i, x_i, q(i)$ and $\dot{\theta}$.

We obtain that this function can be written

$$Q(\theta \mid \dot{\theta}) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \delta(k \mid y_i, x_i, q(i); \dot{\theta}) \log(f(y_i, k \mid x_i, q(i); \theta))$$

where

$$\delta(z \mid y_i, x_i, q(i); \dot{\theta}) = \frac{\dot{\delta}_{z,q(i)} \frac{1}{\dot{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(y_i - x_i'\dot{\beta} - \dot{\gamma}_z)^2}{2\dot{\sigma}^2}\right)}{\sum_{k=1}^K \dot{\delta}_{k,q(i)} \frac{1}{\dot{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(y_i - x_i'\dot{\beta} - \dot{\gamma}_k)^2}{2\dot{\sigma}^2}\right)}.$$

The EM algorithm consists to start with a given value $\dot{\theta} = \theta_0$.

Then to maximize $Q(\theta | \dot{\theta})$ with respect to θ in order to obtain a value θ_1 of the vector of parameters and, then, iterate by replacing $\dot{\theta}$ by θ_1 .

We then stop to iterate when two successive values of θ are close enough.

Remark: It is possible to replace the conditional expectation in the definition of $Q(\theta | \dot{\theta})$ by an empirical average obtained using i.i.d. draws of the unobserved component obtained using the conditional distribution of z given $y_i, x_i, q(i)$ and $\dot{\theta}$.

We obtain in this case a **simulated EM algorithm**.

Let us consider the objective function

$$\hat{Q}(\theta \mid \dot{\theta}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{H} \sum_{h=1}^H \log(f(y_i, z_{ih} \mid x_i, q(i); \theta))$$

where $z_{ih} \in \{1, \dots, K\}$ is a draw from the distribution described by the conditional probabilities $\delta(z \mid y_i, x_i, q(i); \dot{\theta})$.

Each iteration of the simulated EM algorithm consists to obtain **new draws** z_{ih} in order to calculate the objective function $\hat{Q}(\theta | \dot{\theta})$ and, then, **to maximize it with respect to θ** .

This function is proportional to

$$\hat{Q}(\theta | \dot{\theta}) \propto \frac{1}{n} \sum_{i=1}^n \frac{1}{H} \sum_{h=1}^H \left[-\frac{1}{2\sigma^2} (y_i - x_i' \beta - \gamma_{z_{ih}})^2 - \frac{1}{2} \log(\sigma^2) \right]$$

H is the number of draws per transaction. n is the number of transactions.

We have Q districts in the data set. A given transaction occurs in one of these Q districts.

Unobserved heterogeneity specific to district q is denoted z_{ih}^q and let us consider that $z_{ih} \equiv z_{ih}^{q(i)}$.

The variable $z_{i,h;k,q} = \mathbf{1}[z_{ih}^q = k]$ is a binary variable. By convention $z_{i,h;k,q}$ is set to zero if $q \neq q(i)$ where $q(i)$ is the district where takes place transaction i .

$M_{iq} = [z_{i,h;k,q}]_{h=1,\dots,H;k=1,\dots,K}$ is a set of rectangular matrices (with H lines and K columns) and $M = [M_{iq}]_{1 \leq i \leq n; 1 \leq q \leq Q}$ is a rectangular matrix ($n \times H$ lines and $K \times Q$ columns).

The matrix of observed explanatory is $A = [A_i]_{1 \leq i \leq n}$ where $A_i = x_i' \otimes \mathbf{1}_H$ and $\mathbf{1}_H$ is a vector of ones with H elements.

The value of $\alpha = (\beta', \gamma')$ that maximizes $\hat{Q}(\theta \mid \dot{\theta})$ is

$$\hat{\alpha} = (X'X)^{-1} X'Y$$

where $X = (AM)$ and $Y = y \otimes \mathbf{1}_H$, where y is the vector of logarithms of transaction prices.

The **estimator of the variance** of the error terms is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{H} \sum_{h=1}^H \left[(y_i - x_i' \hat{\beta} - \hat{\gamma}_{z_{ih}})^2 \right]$$

The **estimator of the probability** that a transaction that occurs in district q belongs to category z can be alternatively estimated by

$$\hat{\delta}_{z,q} = \frac{\sum_{j \in T_q} \sum_{h=1}^H z_{j,h;z,q}}{n_q \times H}$$

where T_q is the set of indexes of transactions that took place in district q and n_q is the number of these transactions.

At each iteration of the algorithm, to obtain draws of the unobserved heterogeneity components in the conditional distribution, we use uniform random draws on the interval $[0; 1]$.

These uniform draws are re-drawn at each iteration of the EM algorithm, we obtain a **StEM algorithm** (see Nielsen, 2000). The asymptotic variance-covariance matrix associated to the

StEM algorithm is (see Nielsen, 2000)

$$\Sigma(\theta_0) = I(\theta_0)^{-1} + \frac{1}{H} V(\theta_0)^{-1} E_{\theta_0}[I_Y(\theta_0)] V(\theta_0)^{-1} (I - F(\theta_0)^2)^{-1}$$

where θ_0 is the true value of θ and I is the identity matrix.

The matrix $V(\theta)$ can be estimated by

$$\hat{V}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{1}{H} \sum_{h=1}^H \mathbf{s}_{y_{ih}^*}(\theta) \mathbf{s}_{y_{ih}^*}(\theta)'$$

where $y_{ih}^* = (z_{ih}, y_i)$ and the matrix $E_{\theta}[I_Y(\theta)]$ can be estimated by

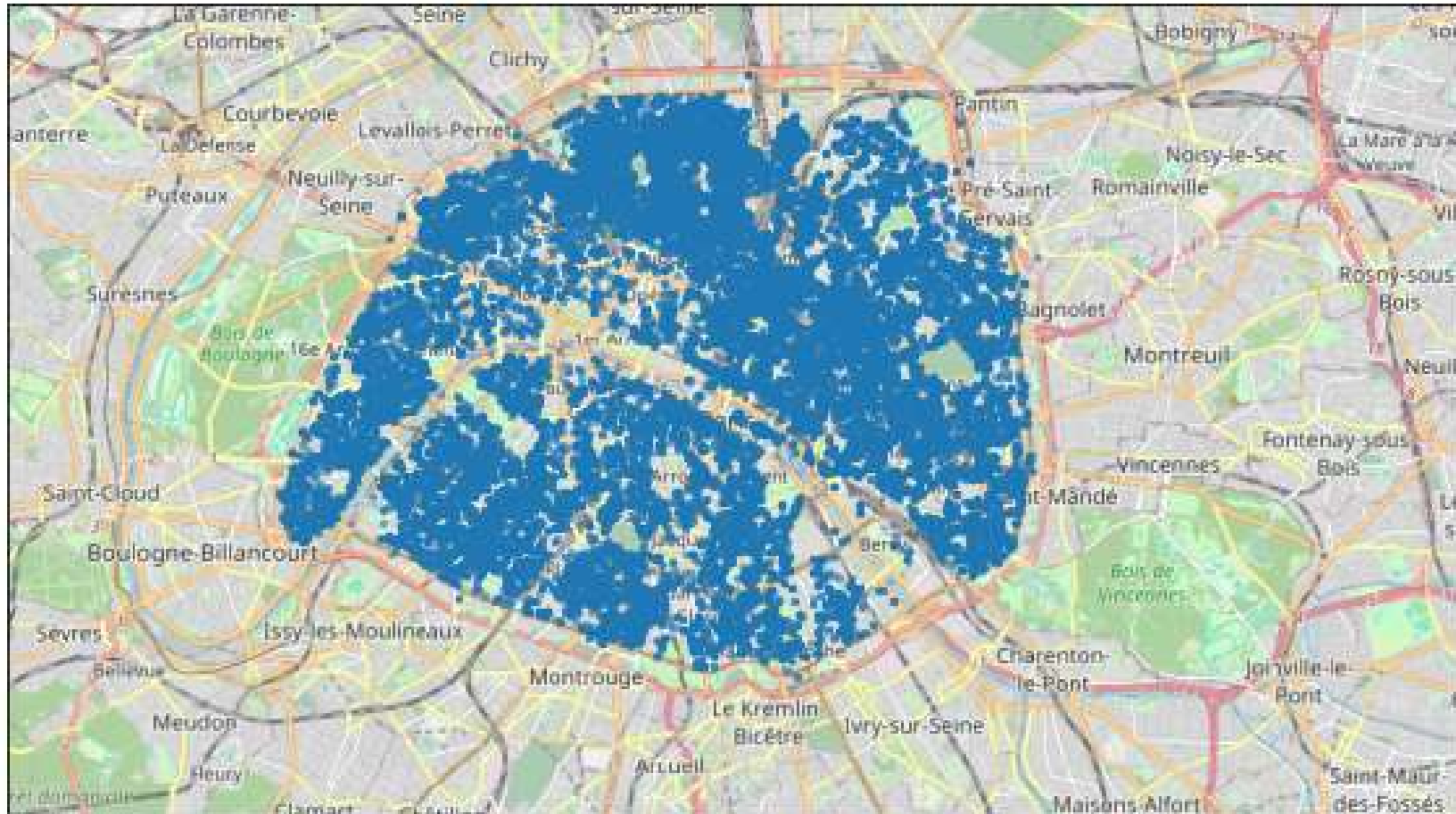
$$\hat{E}_{\theta}[I_Y(\theta)] = \frac{1}{n} \sum_{i=1}^n \frac{1}{H} \sum_{h=1}^H \mathbf{s}_{y_{ih}^*|y_i}(\theta) \mathbf{s}_{y_{ih}^*|y_i}(\theta)'$$

The information matrix

$I(\theta) = E_{\theta}[\mathbf{s}_y(\theta) \mathbf{s}_y(\theta)' | x] = V(\theta) - E_{\theta}[I_Y(\theta)]$. The expected share of missing information is $F(\theta) = E_{\theta}[I_Y(\theta)] V(\theta)^{-1}$

- Administrative data from DVF (Demande de Valeur Foncière) for the period 2018 to 2024.
 - Allow you to know about real estate transactions in mainland France and the French Overseas Territories, with the exception of Alsace, Moselle and Mayotte.
 - Select transactions on apartments in Paris with price per m² between 1000 and 100000 euros.
 - Environmental noise over a full day (road, train, airplane) from Bruitparif. [▶ Noise map](#)
 - Location of primary and secondary education establishments (education.gouv.fr). [▶ 'Best' highschools](#)
 - Transport infrastructure information from RATP and Ile de France Mobilité.
 - Annual rates of residential burglaries and attempted burglaries from Police Nationale and Gendarmerie Nationale (SSMSI) [▶ Burglary rates](#).

Sample



Location of transactions on apartments in Paris (2023).

Table: Sample characteristics

Apartment price (Euros)

Transaction price	517856.31
Price per square meter	9523.76
By district of Paris:	
1 ^{er} arrondissement	12007.44
2 ^e arrondissement	10586.02
3 ^e arrondissement	11157.52
4 ^e arrondissement	12942.19
5 ^e arrondissement	11207.50
6 ^e arrondissement	13545.21
7 ^e arrondissement	13718.71
8 ^e arrondissement	11490.79

Note : DVF 2018 (price per m^2 between 1000 and 100000 euros).

Table: Sample characteristics

9 ^e arrondissement	10040.55
10 ^e arrondissement	9176.41
11 ^e arrondissement	9387.93
12 ^e arrondissement	8551.32
13 ^e arrondissement	8260.10
14 ^e arrondissement	9181.73
15 ^e arrondissement	9091.78
16 ^e arrondissement	10338.74
17 ^e arrondissement	9613.16
18 ^e arrondissement	8375.58
19 ^e arrondissement	7574.39
20 ^e arrondissement	7936.11

Note : DVF 2018 (price per m^2 between 1000 and 100000 euros).

Table: Sample characteristics

Number of rooms	
1 room	28.11
2 rooms	33.72
3 rooms	21.84
4 rooms	9.70
5 rooms	4.12

Note : DVF 2018 (price per m^2 between 1000 and 100000 euros).

Table: Sample characteristics

Transportation	
Distance to center (Place Dauphine)	3431.42
Distance to nearest subway station	237.25
Share of stations outside Paris	8.73
Average traffic	2892171.63
One Connection	0.8258
Two Connections	0.1278
Three Connections	0.0364
Four or five Connections	1.01

Educational facilities (distance in meters)	
Nearest high school	358.31
Nearest best high school	1659.81

Number of individuals	35499
------------------------------	--------------

Note : DVF 2018 (price per m^2 between 1000 and 100000 euros).

Table: Noise Pollution

Noise (dB)	Proportion of transactions
<40	8.53
40-45	35.84
45-50	29.35
50-55	11.24
55-60	6.03
60-65	4.82
65-70	3.23
70-75	0.92
75-80	0.04
>80	0.01

Percentages. Data source : BruitParif and DVF.

Table: Logarithm of the transaction price

Discrete panel data method (Stochastic EM)				
	K=2	K=3	K=4	
	(10)	(11)	(12)	
Unobserved het. values				
γ_1	8.41712*** (0.01637)	7.84387*** (0.01980)	7.56337*** (0.02186)	
γ_2	9.64260*** (0.01107)	8.80008*** (0.01477)	8.02598*** (0.02072)	
γ_3		9.70054*** (0.01364)	8.92013*** (0.01844)	
γ_4			9.72763*** (0.01710)	

Number of draws $H = 100$. Standard errors calculated by bootstrap.

(*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price

Discrete panel data method (Stochastic EM)			
	K=2	K=3	K=4
	(10)	(11)	(12)
2019	0.06692*** (0.00240)	0.06827*** (0.00159)	0.06776*** (0.00158)
2020	0.13874*** (0.00206)	0.13911*** (0.00168)	0.13828*** (0.00210)
2021	0.13247*** (0.00204)	0.12979*** (0.00196)	0.12858*** (0.00198)
2022	0.12269*** (0.00210)	0.11731*** (0.00220)	0.11571*** (0.00247)
2023	0.06677*** (0.00225)	0.06544*** (0.00202)	0.06538*** (0.00242)
2024	0.01287*** (0.00226)	0.01174*** (0.00194)	0.01084*** (0.00245)

Number of draws $H = 100$. Standard errors calculated by bootstrap.

(*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price

Discrete panel data method (Stochastic EM)			
	K=2	K=3	K=4
	(10)	(11)	(12)
Two rooms	0.02567*** (0.00160)	0.02664*** (0.00207)	0.02655*** (0.00182)
Three rooms	0.04663*** (0.00205)	0.04882*** (0.00262)	0.04860*** (0.00251)
Four rooms	0.06276*** (0.00294)	0.06566*** (0.00384)	0.06451*** (0.00366)
Five rooms and more	0.05294*** (0.00553)	0.05316*** (0.00644)	0.05131*** (0.00619)
Log(Surface)	0.87830*** (0.00359)	0.86054*** (0.00471)	0.85236*** (0.00619)
Surface	0.00232*** (0.00008)	0.00258*** (0.00010)	0.00274*** (0.00012)
Dependence	0.02499*** (0.00161)	0.02909*** (0.00152)	0.03036*** (0.00174)
Distance to center (Place Dauphine)	-0.27267*** (0.00741)	-0.25714*** (0.00819)	-0.25872*** (0.00841)

Number of draws $H=100$. Distance divided by 10000. Standard errors calculated by

bootstrap. (*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price

	Discrete panel data method (Stochastic EM)		
	K=2	K=3	K=4
	(10)	(11)	(12)
Distance to station $\leq 200\text{m}$	-0.01393*** (0.00138)	-0.01270*** (0.00125)	-0.01195*** (0.00138)
Distance to station $\geq 300\text{m}$	0.00078 (0.00117)	0.00169* (0.00129)	0.00250** (0.00119)
Nearest station is in Paris:			
Constant	0.01561*** (0.00207)	0.01926*** (0.00207)	0.01855*** (0.00197)
Passenger traffic	-0.00413*** (0.00020)	-0.00410*** (0.00017)	-0.00420*** (0.00019)
Two connections	-0.00626*** (0.00196)	-0.00485** (0.00166)	-0.00488*** (0.00160)
Three connections	-0.01719*** (0.00373)	-0.01255*** (0.00433)	-0.01045*** (0.00353)
Four or five connections	0.00887** (0.00604)	0.00303* (0.00703)	0.00280 (0.00508)

Number of draws $H=100$. Passenger traffic in millions. Standard errors calculated by

bootstrap. (*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price

	Discrete panel data method (Stochastic EM)		
	K=2	K=3	K=4
	(10)	(11)	(12)
Noise Pollution (Road, Train, Plane)			
45-60 dB	-0.00502*** (0.00123)	-0.00566*** (0.00108)	-0.00562*** (0.00108)
>60 dB	-0.00169 (0.00189)	0.00117 (0.00173)	0.00213 (0.00192)
Crime (burglary rate)			
> 13.5	0.13926*** (0.00183)	0.14166*** (0.00220)	0.14475*** (0.00212)
Education			
Distance to high school	-0.58359*** (0.03060)	-0.59273*** (0.02556)	-0.56890*** (0.02842)
Distance to the best High Schools	-0.61781*** (0.00683)	-0.63099*** (0.00569)	-0.63187*** (0.00729)
σ^2	0.06960*** (0.00041)	0.05779*** (0.00034)	0.05437*** (0.00032)
N	232091	232091	232091

(*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price (continuation)

	Discrete panel data method (Stochastic EM)		
	K=2 (10)	K=3 (11)	K=4 (12)
Label since July 2021			
A or B	0.14451* (0.08023)	0.12390 (0.09896)	0.27084** (0.11409)
C	0.02311** (0.00918)	0.02325** (0.00942)	0.02215*** (0.00720)
D	ref.	ref.	ref.
E	-0.01484*** (0.00465)	-0.01534*** (0.00440)	-0.01734*** (0.00474)
F	-0.02041*** (0.00547)	-0.02126*** (0.00485)	-0.02343*** (0.00490)
G	-0.03353*** (0.00494)	-0.03453*** (0.00493)	-0.03368*** (0.00508)

(*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Table: Logarithm of the transaction price (continuation)

	Discrete panel data method (Stochastic EM)		
	K=2	K=3	K=4
	(10)	(11)	(12)
Label before July 2021			
A or B	-0.03768*** (0.00620)	-0.03298*** (0.00460)	-0.03392*** (0.00553)
C	-0.00196 (0.00561)	-0.00440 (0.00490)	-0.00328 (0.00564)
D	-0.00794* (0.00461)	-0.00913** (0.00424)	-0.00944* (0.00461)
E	-0.01745*** (0.00482)	-0.01597*** (0.00431)	-0.01615*** (0.00434)
F	-0.02691*** (0.00530)	-0.02530*** (0.00470)	-0.02610*** (0.00533)
G	-0.02560*** (0.00552)	-0.02655*** (0.00475)	-0.02830*** (0.00517)
Controls	Yes	Yes	Yes
N	93447	93447	93447

(*) Significant at 10%. (**) Significant at 5%. (***) Significant at 1%.

Conclusion

Physical characteristics of apartments (surface, number of rooms, presence of a dependence, energy efficiency) are important determinant of the price.

This is also the case for neighborhood characteristics (education, delinquency, transportation, noise pollution).

The impact of the characteristic of neighborhood depends of the distance to the accommodation (education). This impact can vary with distance (transportation). Can depend on other features (passenger traffic and number of connections for transportation).

Further research may consists in studying the impact of other characteristics of the neighborhood as proximity to the Seine or to prestigious avenues (Champs-Élysées, Foch).

The distribution of neighborhood-specific unobservable heterogeneity (80 districts).

Table: Best High Schools in Paris

Rank	Name	Rank	Name
1	Lycée Henri IV	6	Lycée Fénelon Sainte Marie
2	Lycée Lous-Le-Grand	7	Lycée Saint Michel de Picpus
3	Lycée Stanislas	8	Lycée Les Francs Bourgeois
4	Lycée Janson de Sally	9	Lycée Racine
5	Lycée Claude Monnet	10	Lycée Hélène Boucher

◀ Go Back

