How platform data and machine learning methods can give a better understanding of rent to price ratios' determinants

Martin REGNAUD (*), Marie BREUILLE (*), Julie LE GALLO (*)

(*) CESAER UMR1041, INRAE, L'Institut Agro, Université de Bourgogne Franche-Comté

mregnaud@meilleursagents.com

31st March 2022

Context

Rent to price ratios help actors to take their real estate decisions

- "Should I buy or rent my home ?"
- "What profit should I expect from my real estate portfolio?"
- "Did our policy create an advantage of buying over renting ?"

Individual data are necessary to control the difference between rented and sold apartments (Hill and Syed, 2016)
$\rightarrow$ difficult access to individual data of rents in France is a brake

Research question

How can platform data and machine learning improve our knowledge of spatial heterogeneity of rent to price ratios ?

Objectives of this paper

1. Characterize rent to price ratios heterogeneity at a large scale thanks to a quality adjusted matched dataset in France (2010-2022)
2. Differentiated analysis by "Aires d'attraction des villes" at intra and inter-areas level

In both cases, use of machine learning and SHAP framework for a better understanding of ratios and feature interactions

Preliminary results:

- Average ratio tends to decrease as area density increases
- Strong spatial heterogeneity for ratios between attraction areas in France, not only explained by density of the area
- Area effect on rent to price ratio strongly depends on the number of rooms even for an equivalent area

Literature on Rent to Price ratios

International studies

Hill and Syed (2016), Bracke (2015), Campbell et al (2009)

French studies

- OLL data :
  - Gregoir et al (2012), Trouve (2019)
- Web platform data :
  - Chapelle and Eymeoud (2018)

Contributions

Data coverage and quality contribution

1. First study at France national level
2. Source: (DV3F x Meilleurs Agents) 2010-2022
   - Individual rent data with address level information at national scale
   - Extensive description of housing (structured features, text, pictures) $\rightarrow$ exact matching

Methodological contribution

1. Use of Catboost model $\Rightarrow$ Better bias and interaction management + No prior hyp. on interactions
2. Better explainability thanks to shapley values

Property sales data

DV3F dataset (Source : DGFIP, Cerema)

Exhaustive dataset of property transfers in France between 2010 and June 2020.

- Extensive feature description of the property
- Geolocalisation at the individual parcel level
- Detailed information about buyer/ seller profiles

Sample:
2,373,430 apartments sales.
2010-2020, metropolitan area and French overseas departments and territories except Alsace, Moselle, Mayotte
Features : parcel, area, number of rooms, floor, number of terraces, cellar, garage.

Rental data

Rental ads dataset (Source : MeilleursAgents)

Original dataset of rental ads published on Meilleurs Agents between 2000 and 2020.

- Geolocalisation at the address level
- Precise description of the property (features, text description, pictures)
- Unbiased substitute for survey data (Chapelle et Eymeoud 2018)

Sample :
320'000 apartment ads covering 6000 cities in France (86% of rented housing) [1]

1. INSEE, 2014

Matching Strategy

We perform a quality adjusted matching on :

- Same parcel
- Same floor and same number of rooms
- Maximum discrepancy of 2 sq. meters

We obtain **58'500 matched ratios**

Our average apartement is located on the 2nd floor, has 41 sqm and 2 rooms

Descriptive statistics on features of matched ratios

Variable	Mean (std)	Min	25%	50%	75%	Max
Area	40.79 (19.6)	9	24	38	54	246
Floor	2.29 (1.9)	0	1	2	3	10
Number of rooms	1.96 (1.0)	1	1	2	3	7
Number of terrace	0.1 (0.3)	0	0	0	0	3
Number of cellar	0.46 (0.5)	0	0	0	1	4
Number of garage	0.02 (0.2)	0	0	0	0	3
Elevator_1	0.46 (0.5)	0	0	0	1	1
Elevator_NA	0.18 (0.4)	0	0	0	0	1
Elevator_0	0.37 (0.5)	0	0	0	1	1
Prop. secondary residence	0.05 (0.1)	0	0.01	0.02	0.04	0.84
Housing stock index	28.65 (95.6)	0	1.10	3.52	14.37	2101.53
Prop. apartments	0.79 (0.2)	0.02	0.69	0.88	0.96	1.00
Prop. vacant housings	0.08 (0.04)	0	0.05	0.07	0.1	0.42
Median declared income (k€)	24.5 (7.7)	3.2	19.5	22.6	28.2	66.1
Ratio	0.06 (0.02)	0.01	0.05	0.06	0.07	0.33

Source: DV3F, Meilleurs Agents.

Hedonic explanation of ratios (Rosen 1974)

Differentiated goods have a ratio valued only based on their attributes

$\Rightarrow$ Ratios (r(X)) can be explained as a sum of participations of each feature to the ratio.

$$r(X_i) = \beta_0 + \sum_{j=1}^{N} \beta_j X_{ij} \tag{1}$$

N Number of features describing the properties
r_i = Ratio for property i
X_{ij} = Value of feature j for property i
β_0 = Ratio of the reference property
β_j = Fixed effect of feature j on the ratio value

SHapley Additive exPlanations model (SHAP) (Lundberg et Lee 2017)

Shap values provide a local explanation model g for each property that matches the original model on this particular property

$$r(X_i) = g(X_i') = \phi_0 + \sum_{j=1}^{M} \phi_j X_{ij}' \tag{2}$$

M Number of features in the simplified input space
$r(X_i)$ = Ratio for property i
$g(X_i')$ = Local linear explanation model for property i X_{ij}' = Value of simplified feature j for property i
ϕ_0 = Ratio of the reference property for model g
ϕ_j = Fixed Effect of simplified feature j on the ratio value for model g

$\Rightarrow$ Ratios can be explained as a sum of participations of each feature to the ratio. **Except** $(\phi_j)_{j \in [\![1,M]\!]}$ depend on the property.

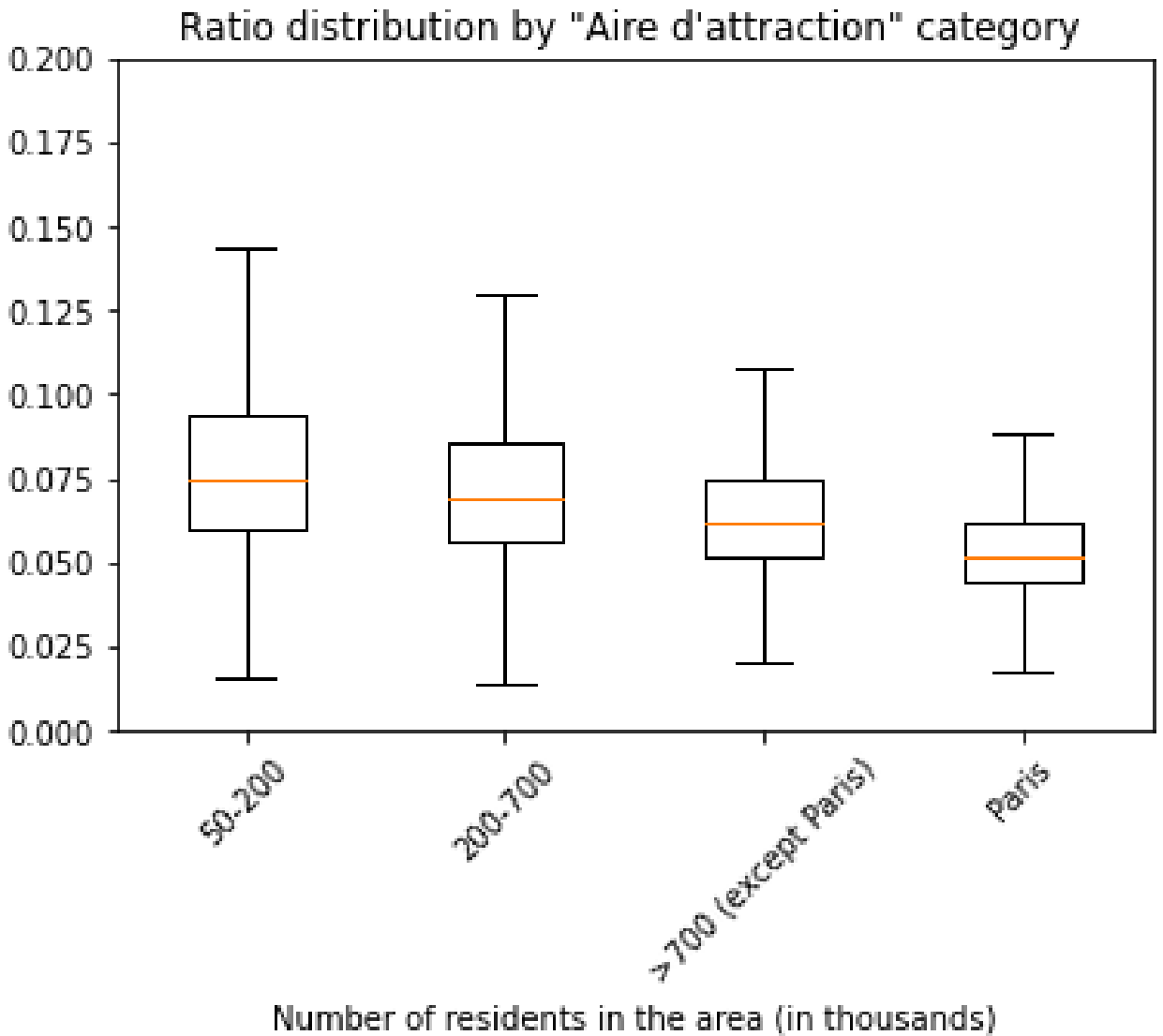
Empirical Strategy

We train a Catboost model (Prokhorenkova et al, 2018) using the following features

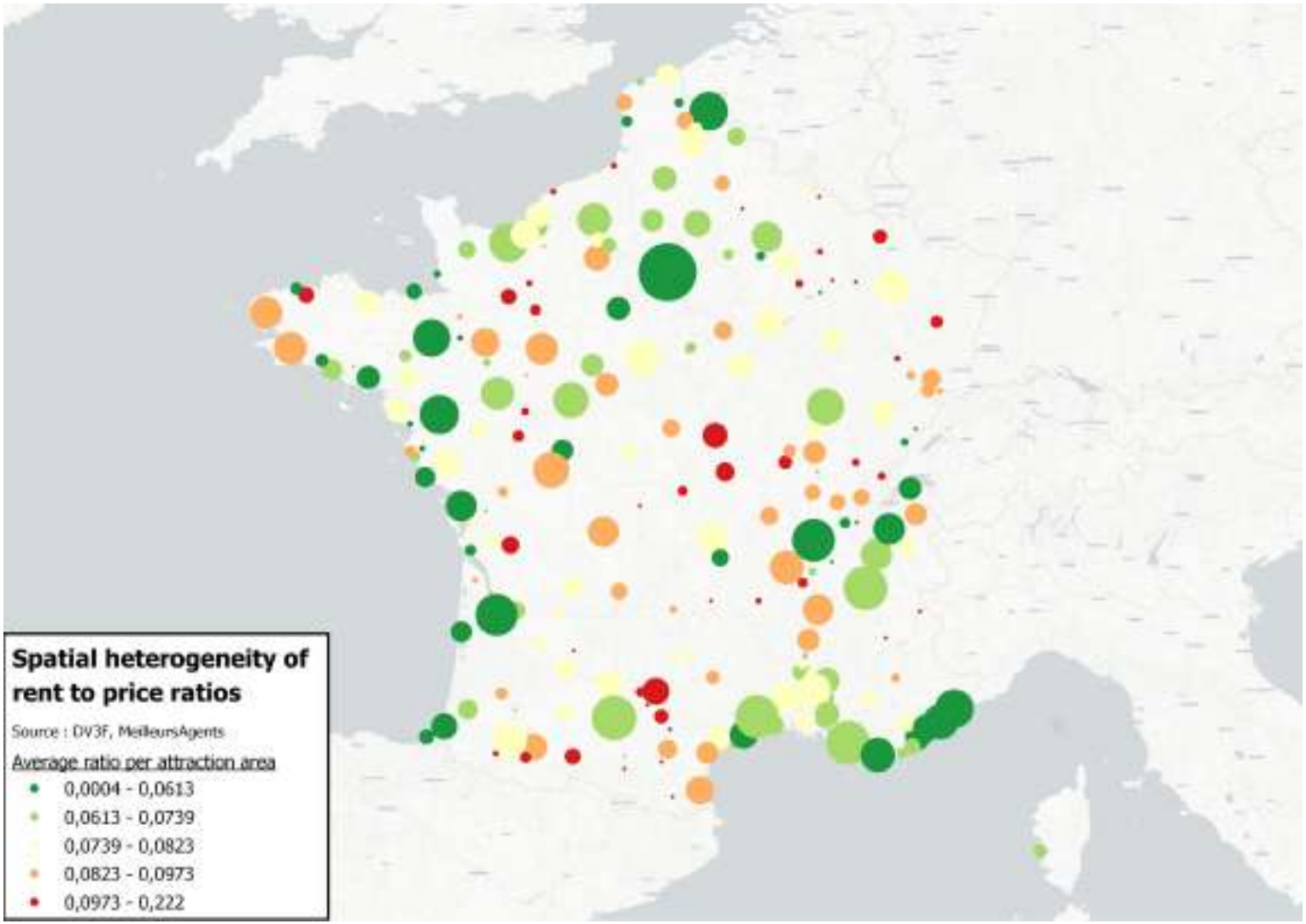
- area
- floor clipped to 10
- number of terraces
- number of cellars
- number of parking lots
- Existence of an elevator
- Furnished rent
- Department ID
- Population category of attraction area
- Proportion of apartments among total housings in the IRIS
- Proportion of vacant housings in the IRIS
- Proportion of secondary residence among housings in the IRIS
- Housing stock youth index
- Median declared income per consumption unit in the IRIS

All housings and income data come from INSEE survey 2014

Average ratios decrease when area size increases



A strong spatial heterogeneity between attraction areas



Results from hedonic model (1/3)

Results of hedonic regression model (R^2 : 0.48)

	Coeff (std)	0.025	0.975
Intercept	0.1044*** (0.001)	0.102	0.107
area	-0.0005*** (8.04e-6)	-0.001	-0.000
furnished	0.0059*** (0.000)	0.006	0.006
rooms_1	0.0004* (0.000)	-0.000	0.001
rooms_3	0.0067*** (0.000)	0.006	0.007
rooms_4	0.0149*** (0.000)	0.014	0.016
rooms_5	0.0262*** (0.001)	0.024	0.028
rooms_6	0.0459*** (0.001)	0.039	0.053
terrace_1	-0.0058*** (0.000)	-0.006	-0.005
terrace_2	-0.0061*** (0.002)	-0.009	-0.003
terrace_3	-0.0189 (0.012)	-0.042	0.004
elevator_1	-0.0009*** (0.000)	-0.001	-0.001
elevator_NA	0.0005** (0.000)	0.000	0.001
Proportion of secondary residence	-0.0233*** (0.001)	-0.025	- 0.021
Housing stock youth index	2.245e-6** (6.48e-7)	9.74e-7	3.52e-6
Prop. of apartments	-0.0053*** (0.000)	-0.006	-0.005
Prop. of vacant housing	0.0495*** (0.002)	0.045	0.053
Median dec. inc. per consumption unit	-5.107e-7*** (1.11e-8)	-5.33e-7	-4.89e-7

*** : $p < 0.001$ / ** : $p < 0.01$ / * : $p < 0.1$

Results from hedonic model (2/3)

Results of hedonic regression model

	Coeff (std)	0.025	0.975
rented_as_new	-0.0006 (0.001)	-0.003	0.002
cellar_1	0.0008*** (0.000)	0.000	0.001
cellar_2	0.0044*** (0.000)	0.004	0.005
cellar_3	0.0051 (0.003)	-0.001	0.011
cellar_4	0.0044 (0.007)	-0.010	-0.019
garage_1	-0.0019*** (0.001)	-0.003	-0.001
garage_2	-0.0086*** (0.001)	-0.011	-0.007
garage_3	-0.0130 (0.008)	-0.029	0.003
floor_0	0.0025*** (0.000)	0.002	0.003
floor_1	-0.0002 (0.000)	-0.000	0.000
floor_3	-0.0001 (0.000)	-0.000	0.000
floor_4	0.0006** (0.000)	0.000	0.001
floor_5	-0.0005 (0.000)	-0.001	0.000
floor_6	0.0005 (0.000)	-0.000	0.001
floor_7	0.0003 (0.001)	-0.001	0.001
floor_8	0.0008 (0.001)	-0.001	0.002
floor_9	0.0004 (0.001)	-0.001	0.002
floor_10_or_more	0.0009 (0.001)	-0.000	0.002

*** : $p < 0.001$ / ** : $p < 0.01$ / * : $p < 0.1$

Results from hedonic model (3/3)

Results of hedonic regression model

	Coeff (std)	0.025	0.975
Outside of cities attraction	-0.0120***(0.003)	-0.018	-0.006
< 10 000 inhab.	-0.0033 (0.003)	-0.009	0.003
10k - 20k inhab.	0.0116*** (0.001)	0.009	0.014
20k - 30k inhab.	-0.0035* (0.002)	-0.007	-0.000
30k - 50k inhab.	-0.0002 (0.001)	-0.002	0.001
50k - 75k inhab.	0.0023**(0.001)	0.001	0.004
75k - 100k inhab.	-0.0006 (0.001)	-0.003	0.002
100k - 125k inhab.	-0.0004 (0.001)	-0.003	0.002
125k - 150k inhab.	-0.0029* (0.001)	-0.005	-0.001
150k - 200k inhab.	0.0005 (0.001)	-0.002	0.003
200k - 300k inhab.	-0.0062*** (0.001)	-0.008	0.005
300k - 400k inhab.	-0.0015* (0.001)	-0.003	0.000
400k - 500k inhab.	-0.0079*** (0.001)	-0.010	-0.006
500k - 700k inhab.	-0.0073*** (0.001)	-0.009	-0.005
>1M inhabitants (outside Paris)	-0.0034*** (0.001)	-0.005	0.002
Paris area	-0.0066 (0.002)	-0.010	-0.003

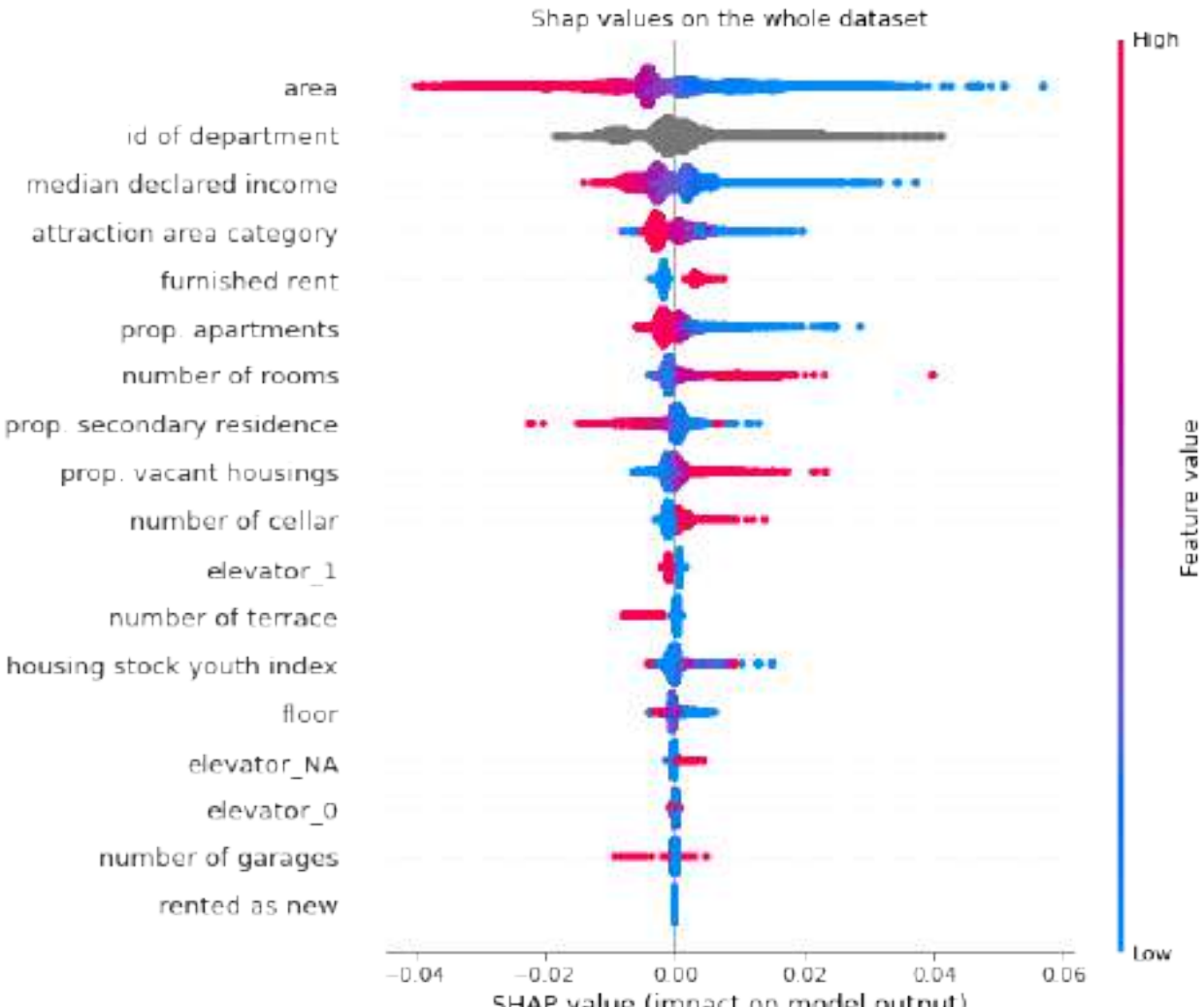
*** : $p < 0.001$ / ** : $p < 0.01$ / * : $p < 0.1$

Shapley values explain the effect of each feature

How to read :

- The fact that this particular apartment **has 73 sqm decreases its ratio by 0.02** compared to the reference apartment (*base_value*).
- Yet the fact that it is located in an **area having between 30k and 50k inhabitants** (*area_category* = *14*) **increases it by 0.006** compared to the reference

Our shapley explanation model is consistent with our hedonic model with better variance explanation ($R^2 = 0.58$)



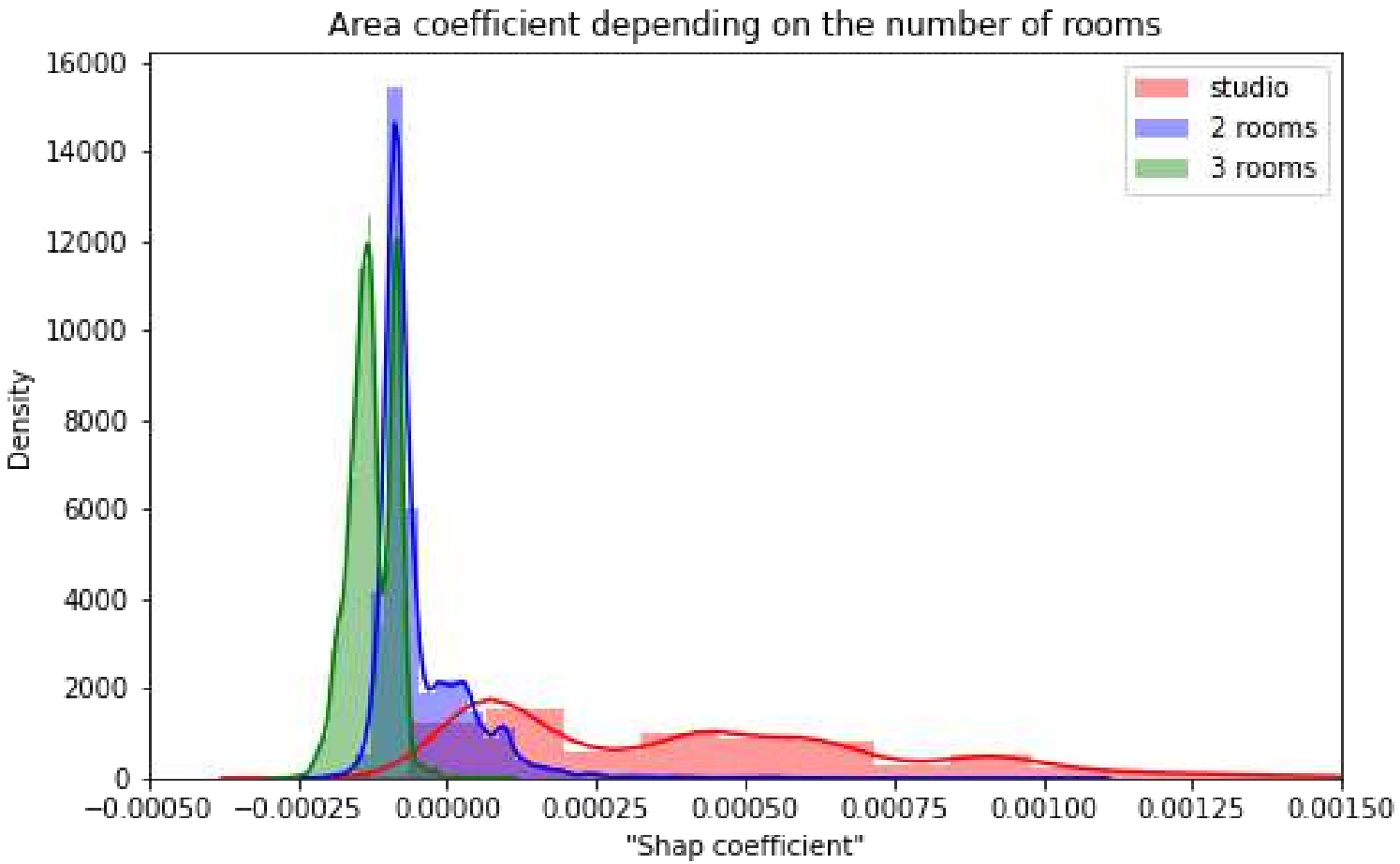
The area coefficient varies significantly

We isolate ϕ_{area} by dividing the shap value effect ($\phi_{area} X_{area}$) by the value of the area.

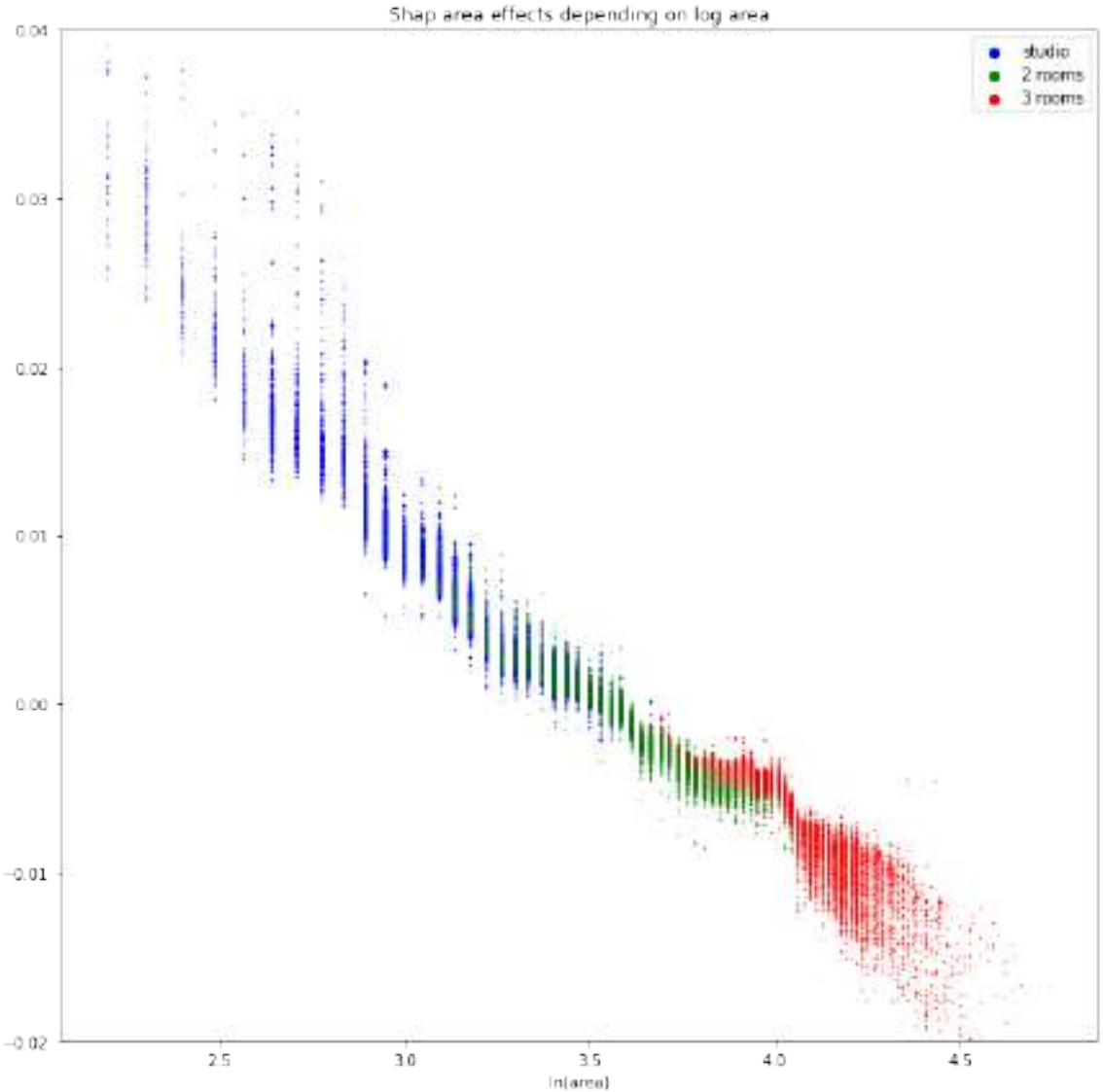
Variable	**Mean (std)**	**25%**	**50%**	**75%**
area ($\times 10^{-4}$)	1.25 (4.11)	-0.101	-0.53	2.14

Table – Area "Shap coefficients" distribution in our model for all France

The area "shap coefficient" decreases with the number of rooms



For a 45 sqm area, three rooms yield a higher ratio than two rooms



Conclusion

1. There exists geographic disparities of rent to price ratios between attraction areas :
   5.4% in Paris / 9.01% in Limoges

2. Shap value methods can be used to better understand ratios mechanisms :
   Hedonic regression model : $R^2 = 0.48$
   Catboost model $R^2 = 0.58$

Next

1. Apply shapley values approach to all features
2. Robustness checks
3. Is intra area spatial heterogeneity the same across areas ?
4. Do rent to price ratios adjust to reach an equilibrium position ? Are the fundamentals of this position the same in all areas ?
5. How do this heterogeneity adjust across time ? Did Covid-19 crisis affected attraction areas differently ?

$\Rightarrow$ Causal inference needed

Thank you!

Intra Paris spatial heterogeneity