

Rapport de stage

Difference-in-Discontinuities: a Method for Territorial Policy Evaluation A Case Study on Rent Control in Paris

Master
Statistiques
et
Econométrie

Année scolaire
2023 - 2024

André Miranda Acevedo

Yacine Allam
Professional Tutor
Avivacity School for Technology, Business & Society

Nathalie Picard
Marie Boltz
Academic Tutors
Faculté des Sciences Économiques et de Gestion
Université de Strasbourg



Acknowledgements

Six months ago, I joined **aivancity: School for Technology for Business & Society** to complete my final internship. I learned plenty about myself and my field of studies. I would like to thank Christophe de Beauvois and Tawhid Chtioui for giving me this opportunity. In particular, I express my gratitude to Yacine Allam for being my mentor and showing me the way forward, and to the entire staff of **aivancity** for welcoming me as part of the team.

I also want to thank the academic team of the **Master Statistiques Econométrie** for my education and all their teachings over the years.

Finally, I thank all the people who have been by my side during these five years, my family and my friends.

Master Statistiques et Econométrie

Rapport de stage

Jury

Marie Boltz
Nathalie Picard
Patrick Rondé

Année scolaire
2023 - 2024



Contact information

andre9miranda@gmail.com
33.7.55.64.87.10

Contents

1 Preface	3
2 Introduction	4
3 Literature Review	6
3.1 Causal Inference Methods	6
3.2 Optimal Bandwidth Choice	8
4 Methods	11
4.1 Differences-in-Differences	12
4.1.1 Setup	12
4.1.2 Identification	13
4.1.3 Estimation	14
4.2 Regression Discontinuity Design	15
4.2.1 Setup	15
4.2.2 Identification	15
4.2.3 Estimation	17
4.3 Difference-in-Discontinuities	17
4.3.1 Setup	18
4.3.2 Identification	18
5 Diff-in-Disc Estimation	22
5.1 Local Linear Regression	22
5.2 Optimal Bandwidth	22
5.2.1 Plug-in Estimator (Imbens & Kalyanaraman, 2012)	22
5.2.2 Robust Bias-Corrected Estimator (Calonico et al., 2014)	23
5.2.3 Adaptive Stochastic Gradient Descent	23
6 Simulation	28
6.1 Data	28
6.2 Results	30
6.2.1 Biased Results for RD and DiD	30

6.2.2	Diff-in-Disc results for different Bandwidth Estimators	31
7	Application - Estimating the Effect of Rent Control in Paris on Real Estate Prices	32
7.1	Context	32
7.2	Data	34
7.3	Diff-in-Disc Analysis	45
8	Conclusion	50

1 Preface

This memoir recounts my experience as an intern at aivancity School for Technology, Business & Society. This institution stands out not only in France, but also across Europe for its unique hybrid approach, blending technology, business, and ethics. Artificial intelligence is not merely a technical subject; it must be understood within its business, societal, and ethical contexts. Based in Paris-Cachan and founded in 2020, the school has grown rapidly. For me, it was an important experience, not only academically but also in terms of interaction with colleagues, professors, and students.

During my internship, I joined a research team working on an article to be published in an economic journal to be defined led by Dr. Yacine Allam, Associate Professor and head of the M.Sc. Data Management. The six-month period was just enough to contribute meaningfully to the project, which continues to evolve with my contributions now incorporated.

I have written this report in English, as the article will be presented in this language, and because my work throughout this process has been conducted in English. I have used LaTeX as it is the collaborative platform that the team works on. LaTeX is the de facto standard for the communication and publication of scientific documents.

2 Introduction

The ability to recognize causal relationships is intrinsic to economic research and is essential in policy evaluation. It is at the heart of econometrics to observe correlations and establish credible cause-and-effect relationships that inform strategic decisions in a world increasingly dominated by data.

Causal inference methods provide the tools necessary to make these distinctions, allowing researchers to effectively isolate the impact of specific interventions or treatments from other influencing factors. This is particularly crucial in evaluating public policies, where isolating the true effects of an intervention can lead to better decision-making and resource allocation.

Basic econometric methods often fall short in dealing with unobserved confounders and the complexities of real-world data, especially when an identification framework is not extensively and rigorously developed to ensure that the applied methodology correctly isolates the treatment effect. The advancement of causal inference methods, such as Difference-in-Differences (DiD) and Regression Discontinuity (RD) constitute two of the most popular designs in the modern literature. By defining and testing the assumptions of causal relationships, these methods offer an exhaustive, rigorous, and robust framework for policy analysis.

Along this paper, I engage in the exploration of these two particular designs and contribute to the ongoing discourse by formalizing and extending the Difference-in-Discontinuities (Diff-in-Disc) methodology, providing a framework that enhances the reliability of causal estimates in policy evaluation, especially in cases where classic RD and DiD assumptions may not hold.

This memoir is part of an exploratory study to develop and formalize a robust causal inference methodology. The motivation behind this work stems from the need for unbiased and reliable methods in the evaluation of territorial public policies. The methodology being constructed is akin to a geographic Regression Discontinuity Design but incorporates a temporal dimension. Specifically, the approach seeks to isolate and estimate the treatment effect by conducting a locally linear regression in the proximity of a threshold of the explanatory variable.

For this purpose, this paper begins by first reviewing the literature on Regression Discontinuity (RD) and Difference-in-Discontinuities (Diff-in-Disc) designs, followed by a discussion of the estimation techniques for both RD and Diff-in-Disc, which primarily revolve around non-parametric regression techniques. This section emphasizes the optimal bandwidth choice from non-parametric literature. I contribute to this literature by proposing a data-driven Stochastic Gradient Descent (SGD) to optimize the bandwidth choice.

The following section delves into the theoretical frameworks of the Difference-in-Differences (DiD), RD, and Diff-in-Disc designs, as well as the estimation methods used in each. In this section, I propose an alternative identification assumption for the Diff-in-Disc method. This is followed by an in-detailed description of the adaptive SGD algorithm used for this research.

To complement this theoretical framework, these sections are followed by a study using simulated data. This contributes to the Diff-in-Disc literature by empirically demonstrating that traditional RD and DiD methods, as well as non-local linear regression, produce biased results when applied without careful consideration, particularly in contexts where there is a confounding policy at the threshold.

This is then followed by an application of the Diff-in-Disc method to real-world data exploring the Difference-in-Discontinuities (Diff-in-Disc) methodology in evaluating the impact of rent control policies on property prices within Paris and its adjacent surroundings, known as the *Petite Ceinture*. I aimed to rigorously estimate the local average treatment effect by employing various bandwidth selection methods and kernels. This memoir then concludes with a reflection on the Diff-in-Disc design and discusses possible extensions of the model as well as current limitations.

3 Literature Review

3.1 Causal Inference Methods

In econometrics, causal inference methods applied to policy evaluation have developed a very extensive body of literature. Authors covered comprehensively the body of work regarding the classical causal inference methods for program evaluation (Abadie & Cattaneo, 2018). In recent times, the econometric literature in policy evaluation has improved on the empirical side thanks to the growing access to data and gaining in rigor by emphasizing identification. Isolation of the treatment effect in causal inference consists of isolating the treatment effect in a particular setup through the development of certain assumptions specific to the design. A very common identification problem in many policy evaluation methods is the lack of control of unobserved confounders.

The Regression Discontinuity Design (RDD) was first introduced by Thistlewaite and Campbell (1960) as a reliable alternative to randomized experiments for social program and intervention evaluation. The RD literature has explored numerous parts of the design such as solving and formalizing identification strategies (Hahn et al., 2001), optimal estimation (Porter, 2003), validity test of the design's assumptions and falsification (McCrary, 2008), and conditioning on covariates (Frölich, 2007). Exhaustive and comprehensive guides have also been published as practitioner guides to Regression Discontinuity Designs including Imbens and Lemieux (2008), Van der Klaauw (2008), Lee and Lemieux (2010) and most recently Cattaneo et al. (2020).

In the context of longitudinal panel data, the RD literature has used different approaches, for example, the first-difference RD estimator controlling for unobserved heterogeneity in Lemieux and Milligan (2008), or the dynamic RD design to identify the dynamic treatment effects in Cellini et al. (2010). In these setups, the particularity rests in identifying the assignment of treatment over time.

In the branch of longitudinal data Regression Discontinuity Designs, “we take a look at RD studies where identification rests on the difference between two cross-

sectional estimators”¹. In Lalive (2008) there is an age threshold and a border threshold, and there are discontinuities across firms and over time in Leonardi and Pica (2013). These two papers are some of the first to use this type of design in the RD literature.

The first study to employ the term “Difference-in-Discontinuities” more commonly known as “Diff-in-Disc” was Grembi et al. (2016), extending the precise identification assumptions for this design. The Diff-in-Disc method was deployed to control for a confounding policy implemented at the same threshold as the policy of interest. From this paper, the Diff-in-Disc methodology was born. In it, there are generally two identification strategies following the general RD literature:

- Continuity assumptions are respected on the period previous to the treatment and posterior to it (Baskaran & Lopes da Fonseca, 2016; Lippmann, 2018)
- Consider treatment assignment at the threshold “as good as random” as in “Close race elections” (Bhalotra et al., 2018; Chicoine, 2017)

The bandwidth choice in Diff-in-Disc analysis is a critical factor in the estimation process of treatment effects and determines the range of data points from the threshold included in the regression, effectively balancing bias and variance in the estimation process. A smaller bandwidth focuses on data points closer to the cut-off, which can reduce bias by ensuring that the observations are more comparable. However, this may increase variance due to the smaller sample size. Inversely, a larger bandwidth includes more data points, potentially reducing variance but at the cost of introducing bias, as the observations may differ more significantly from those at the threshold. Thus, selecting an optimal bandwidth is essential to achieving a precise and unbiased estimate of the Local Average Treatment Effect. The following subsection goes over the literature on this topic choice and presents the method introduced by this paper.

¹Grembi, V., Nannicini, T., & Troiano, U. (2016). “Do fiscal rules matter?” *American Economic Journal: Applied Economics*, 8(3), p. 10

3.2 Optimal Bandwidth Choice

The Diff-in-Disc Design aims to determine the Local Average Treatment Effect on the Treated units at a threshold level ‘ c ’, which is determined by the running variable ‘ X ’, and at a starting date t_0 for treatment based on a time variable t . This paper presents several approaches to obtain an optimal bandwidth for Diff-in-Disc estimation.

Research on bandwidth choice for Regression Discontinuity Designs mainly derives from the non-parametric regression literature. Today, most approaches for bandwidth choice revolve around either plug-in or cross-validation methods from this literature (Fan & Gijbels, 1992; Härdle, 1992; Wand & Jones, 1994). These methods typically rely on objective functions that evaluate the performance of the regression function estimator across its entire support.

For local estimators, a reduced bandwidth can lead to better results. Therefore, data-based procedures have been introduced for local bandwidth selection for kernel estimation of a regression function at a point, yielding consistent and asymptotically normal estimators of the minimum mean square error estimation (Brockmann et al., 1993; Schucany, 1995).

Originally, very few papers in the RD literature focused on this topic (DesJardins & McCall, 2008; Ludwig & Miller, 2007). Ludwig and Miller, 2007 proposes the only cross-validation bandwidth selection procedure for RDD setups. DesJardins and McCall, 2008 uses bandwidth estimators that focus separately on the limits of the regression function below the cutoff and above it rather than on the difference in the limits using different bandwidths on the left and right of the cutoff and an Epanechnikov kernel instead of the triangular (or Edge) kernel.

Imbens and Kalyanaraman (2012) introduces a bandwidth estimator based on an Asymptotic Expansion of the MSE. This data-driven estimator suggests the employment of consistent estimators for the various components of the optimal bandwidth. This optimal bandwidth estimator also includes regularization terms to construct a feasible estimator of the optimal bandwidth for Regression Discontinuity Designs. In Imbens and Kalyanaraman (2012) the estimator has very desirable properties, such as being consistent regardless of the curvature at the cutoff, convergent, and

optimal as cited in Li (1987), meaning that the estimator of the optimal bandwidth converges in probability towards the minimizer of the Mean Squared Error (MSE) of the regression,

$$\frac{\text{MSE}(\hat{h})}{\inf_h \text{MSE}(h)} \xrightarrow{p} 1$$

under the following conditions²:

1. $\forall i \in \{1, \dots, N\}$ (Y_i , X_i) are independent and identically distributed (i.i.d.).
2. The marginal distribution of the forcing variable is continuous and bounded away from 0 at the threshold.
3. The conditional mean is at least three times-continuously-differentiable in an open neighborhood of the threshold.
4. The chosen kernel is non-negative, bounded, differs from zero, and is continuous on a compact interval.
5. The conditional variance is bounded and left and right continuous in an open neighborhood of the threshold.

However, Calonico et al. (2014) states that the MSE-optimal bandwidth estimator produces data-driven confidence intervals that may be biased. They derive new optimal choices for the auxiliary bandwidth required for Robust Bias Corrected (RBC) inference to estimate the bias of the main bandwidth estimator and its variance to calculate the CE-optimal bandwidth choice, which minimizes the coverage error of the interval estimator. Their results are coded in the package Calonico et al. (2015), which is now widely used as the standard for Regression Discontinuity Design applied studies.

I contribute to the optimal bandwidth choice literature for RD design by implementing an Adaptive Stochastic Gradient Descent (SGD) algorithm, frequently used in Machine Learning, to solve non-convex optimization problems. The SGD algorithm runs through each interval of the partitioned data set to improve convergence chances at lower values, finding the bandwidth that minimizes MSE. This

²Li, K.-C. (1987). “Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set”. *The Annals of Statistics*, 15(3), p. 961.

SGD algorithm uses an AdaGrad learning rate. In addition, the MSE is weighed by a penalty term that punishes very small bandwidth choices. In this context, there is no analytical form of the MSE function or its gradient. Therefore, the MSE gradient is evaluated for a bandwidth value ' h_j ' by computing numerical derivatives using a variation of the Finite Difference Method for gradient calculation. The SGD method differs from traditional bandwidth choice methods by being purely data-driven.

4 Methods

Let Y be the variable of interest and W the treatment variable. I use potential outcomes to define causal parameters as in Rubin (1974). Y_w represents the value of the potential outcome of the interest variable based on the value w of the treatment variable W . In the simplest case, W is a binary variable, meaning that there are two possibilities for treatment: the unit is either in the treatment group or it is not. Therefore, there are two possible values for Y_w :

- the unit belongs in the treatment group: $w = 1 \Rightarrow Y_w = Y_1$
- the unit belongs in the control group: $w = 0 \Rightarrow Y_w = Y_0$

This setup implies that the realized outcome for each particular unit depends only on the value of its corresponding treatment and does not rely on the outcome or treatment of any of the other units. This is known as the Stable Unit Treatment Value Assumption (SUTVA).

Assumption 1 (SUTVA). *Potential outcomes for each unit i are unrelated to the treatment status of other individuals.*

$$Y_{wi} = Y_i \Leftrightarrow W_i = w$$

Another fundamental assumption in Causal Inference is the unconfounded treatment assignment.

Assumption 2 (Unconfounded Treatment Assignment). *The potential outcomes are independent of the treatment assignment.*

$$(Y_1, Y_0) \perp\!\!\!\perp W$$

This implies the absence of selection bias, ensuring that the estimated Average Treatment Effect on the Treated (ATET) reflects the true causal effect. Additionally, I assume consistency in the observed outcomes.

Assumption 3 (Consistency). *The observed outcome for any unit under the treatment actually received is equal to the potential outcome for that unit under that treatment condition.*

$$W_i = w \Rightarrow Y_i = Y_i(w), \forall i$$

The observed outcome $Y = WY_1 + (1 - W)Y_0$ reveals whether the unit is treated or not. However, obtaining the treatment effect for each unit $Y_{1i} - Y_{0i}$ is simply not possible because a unit cannot be both treated and untreated. This is known as the Fundamental Problem of Causal Inference. To address this, several methodologies are employed, each relying on different identification strategies to estimate the ATE or ATET.

4.1 Differences-in-Differences

A very commonly used method in the literature, in a longitudinal data setting, is the Difference-in-Differences design, also known as Diff-in-diff (abbreviated DiD). Diff-in-diff solves this identification problem by controlling for how unobserved confounders affect the outcome variable over time. For treatment identification in a Diff-in-diff setup, we have the following key identification assumption: Parallel Trends, meaning that the average outcome would've evolved in the same way in both groups if treatment never happened. While the difference-in-differences (Diff-in-diff) design is a widely used approach in causal inference, it may not always be applicable. This is because the assumption of parallel trends is often unrealistic, as units below and above the cutoff point may not share common tendencies or exhibit identical evolution over time in the hypothetical scenario of no treatment.

4.1.1 Setup

Assume the simplest scenario for Diff-in-Diff, that is two time periods: pre-treatment ($t = 0$), and post-treatment ($t = 1$). We have for Y_{it} and W_{it} the variable of interest and the treatment variable for the unit i at the period t . Let Y_{wit} denote the potential outcome of the interest variable of unit i at time t . To keep consistency with notation, w denotes the group that unit i belongs to, where $w \in \{0, 1\}$. The interest and treatment variables are defined as:

- $W_{it} = \mathbb{1}(t = 1)\mathbb{1}(w = 1)$, where $\mathbb{1}(\cdot)$ is the indicator function, meaning that only if the condition A is met, $\mathbb{1}(A) = 1$
- $Y_{it} = Y_{1it}W_{it} + Y_{0it}(1 - W_{it})$

4.1.2 Identification

As mentioned in the introduction, the particularity of the Diff-in-Diff design is that it allows for control of unobserved confounders that might influence the outcome of Y .

Y_{it} is defined as:

$$Y_{it} = W_{it}\tau_{it} + \mu_i + \delta_t + \epsilon_{it} \quad (1)$$

Where:

- Only Y_{it} and W_{it} are observed
- μ_i denotes the fixed effect for the unit i ,
- δ_t is the time effect at the period t
- ϵ_{it} is the error term such that $\mathbb{E}[\epsilon_{it}|W_{it}] = \mathbb{E}[\epsilon_{it}]$ or simply: $\mathbb{E}[\Delta\epsilon_i|W_{it}] = \mathbb{E}[\Delta\epsilon_i]$ where $\Delta\epsilon_i = \epsilon_{i,t=1} - \epsilon_{i,t=0}$

The two potential outcomes for Y_{it} can be written as:

$$Y_{1it} = \tau_{it} + \mu_i + \delta_t + \epsilon_{it} \text{ for the treatment group}$$

$$Y_{0it} = \mu_i + \delta_t + \epsilon_{it} \text{ for the control group}$$

The key identifying assumption for Diff-in-Diff is the Parallel Trends Assumption.

Assumption 4 (Parallel Trends). *The evolution of the average outcome in the hypothetical case that treatment never happened is equal in both groups. Formally, Parallel Trends imply that the difference between δ_1 and δ_0 is the same in both the control group and treatment group.*

Let $\Delta\delta = \delta_1 - \delta_0$:

$$\Delta\delta_{w=1} = \Delta\delta_{w=0} = \Delta\delta$$

Theorem 4.1 (Diff-in-Diff Identification). *The ATET is identified as follows:*

$$\tau_{ATET} = \mathbb{E}[\tau_{it}|W_{it} = 1] = \mathbb{E}[\Delta Y_{i1}|W_{it} = 1] - \mathbb{E}[\Delta Y_{i0}|W_{it} = 0] \quad (2)$$

Where $\Delta Y_i = Y_{i1} - Y_{i0}$

Proof. the expectation of Y_{it} conditional to the treatment status in each group can be summarized in:

$$\mathbb{E}[Y_{i,1}|W_{i1} = 1] = \mathbb{E}[\tau_{i,1}|W_{i,1} = 1] + \mathbb{E}[\mu_i|W_{i,1} = 1] + \mathbb{E}[\varepsilon_{i1}] + \delta_1$$

$$\mathbb{E}[Y_{i,0}|W_{i1} = 1] = \mathbb{E}[\mu_i|W_{i,0} = 1] + \mathbb{E}[\varepsilon_{i0}] + \delta_0$$

$$\mathbb{E}[Y_{i,1}|W_{i1} = 0] = \mathbb{E}[\mu_i|W_{i,1} = 0] + \mathbb{E}[\varepsilon_{i1}] + \delta_1$$

$$\mathbb{E}[Y_{i,0}|W_{i1} = 0] = \mathbb{E}[\mu_i|W_{i,0} = 0] + \mathbb{E}[\varepsilon_{i0}] + \delta_0$$

The first difference in the treatment group yields:

$$\begin{aligned}\mathbb{E}[\Delta Y_i|W_{i1} = 1] &= \mathbb{E}[Y_{i,1}|W_{i1} = 1] - \mathbb{E}[Y_{i,0}|W_{i1} = 1] \\ &= \mathbb{E}[\tau_{i,1}|W_{i,1} = 1] + \mathbb{E}[\varepsilon_{i1}] - \mathbb{E}[\varepsilon_{i0}] + \Delta\delta\end{aligned}$$

For the first difference in the control group, we have:

$$\begin{aligned}\mathbb{E}[\Delta Y_i|W_{i1} = 0] &= \mathbb{E}[Y_{i,1}|W_{i1} = 0] - \mathbb{E}[Y_{i,0}|W_{i1} = 0] \\ &= \mathbb{E}[\varepsilon_{i1}] - \mathbb{E}[\varepsilon_{i0}] + \Delta\delta\end{aligned}$$

The second difference yields:

$$\mathbb{E}[\Delta Y_i|W_{i1} = 1] - \mathbb{E}[\Delta Y_i|W_{i1} = 0] = \mathbb{E}[\tau_{i,1}|W_{i,1} = 1] = \tau_{ATET}$$

□

4.1.3 Estimation

To estimate the ATET in a Diff-in-Diff setup, replace the expectation expressions in the identification equation with the sample averages. In the case of longitudinal data, the ATET is estimated with the following regression equation:

$$y_{it} = \beta_0 + \beta_1 \mathbb{1}_{\{w=1\}} + \beta_2 \mathbb{1}_{\{t=1\}} + \tau_{it} W_{it} + \varepsilon_{it} \quad (3)$$

Where $\hat{\tau}_{ATET}$ is the OLS estimator of the coefficient accompanying the treatment variable, also written as the interaction term $W_{it} = \mathbb{1}_{\{t=1\}} \mathbb{1}_{\{w=1\}}$.

4.2 Regression Discontinuity Design

The Regression Discontinuity (RD) design is a Causal Inference evaluation design that focuses on identifying the treatment effect in a setup where treatment is assigned only to units having a certain score (index, running variable) above a certain threshold (cutoff). In a Sharp Regression Discontinuity, the identification assumption is that all variables, except for the treatment and outcome, are continuous at the cutoff thus isolating the treatment effect. However, in a setting with unobserved confounders that are also cutoff discontinuous, the RD identification is biased as it would capture not only the treatment effect but also the effect of the cutoff discontinuous confounder in the outcome variable.

4.2.1 Setup

Consider a canonical Regression Discontinuity Design also known as a Sharp Design. The necessary elements for a canonical RDD are the following:

- the score variable X is continuously distributed
- $\exists! c, \forall X_i \geq c, i$ is treated
- Perfect compliance of treatment, i.e., treatment assignment and treatment status are identical

We have the following model:

$$Y_i = f(X_i) + \tau_i W_i + \varepsilon_i \quad (4)$$

Where $W_i = \mathbb{1}_{\{X_i \geq c\}}$ is the treatment variable and f summarizes location-specific characteristics that affect outcomes.

4.2.2 Identification

The idea behind identification in a Sharp RDD is that the treatment assignment mechanism, near the cutoff, resembles the assignment that can be seen in a randomized experiment, so treatment assignment near the cutoff is considered “as good as random”. In this setup, the Sharp RD treatment effect is formally defined as

$$\tau_{RD} \equiv \mathbb{E}[Y_{1i} - Y_{0i} | X_i = c]$$

Assumption 5 (Cutoff Continuity). *Identification of the treatment effect relies on the assumption that all determinants of Y_i except for the treatment are continuous at the threshold, i.e. the effects of the running variable and other potentially unobserved individual-specific factors on the outcome are continuous at the cut-off. In other terms:*

- i. *The regression functions $\mathbb{E}(Y_{1i}|X_i = x)$ and $\mathbb{E}(Y_{0i}|X_i = x)$ are continuous at $x = c$.*
- ii. *$\mathbb{E}(\varepsilon_i|X_i = x)$ is continuous at $x = c$.*

Theorem 4.2 (RD Identification). *Under continuity assumptions (A3), the Local Average Treatment Effect can be identified as:*

$$\tau_{RD} \equiv \mathbb{E}[Y_{1i} - Y_{0i}|X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] \quad (5)$$

With $\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \rightarrow c^+} \mathbb{E}[Y_i|X_i = x]$

and $\lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \rightarrow c^-} \mathbb{E}[Y_i|X_i = x]$

Proof. Let $y^+ = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x]$ and $y^- = \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]$ respectively.

For units above the cutoff:

$$\begin{aligned} y^+ &= \lim_{x \rightarrow c^+} \mathbb{E}[Y_i|X_i = x] \\ &= \lim_{x \rightarrow c^+} \mathbb{E}[f(X) + \tau_i W_i + \epsilon_i] \\ &= \lim_{x \rightarrow c^+} \mathbb{E}[f(X)|X = x] + \lim_{x \rightarrow c^+} \mathbb{E}[\tau_i W_i|X = x] + \lim_{x \rightarrow c^+} \mathbb{E}[\epsilon_i|X = x] \text{ by linearity of expectation} \\ &= \lim_{x \rightarrow c^+} f(X) + \tau_{ATE} + \lim_{x \rightarrow c^+} \mathbb{E}[\epsilon_i|X = x] \end{aligned}$$

Analogously, for units below the cutoff:

$$\begin{aligned} y^- &= \lim_{x \rightarrow c^-} \mathbb{E}[Y_i|X_i = x] \\ &= \lim_{x \rightarrow c^-} \mathbb{E}[f(X) + \epsilon_i] \\ &= \lim_{x \rightarrow c^-} f(X) + \lim_{x \rightarrow c^-} \mathbb{E}[\epsilon_i|X = x] \end{aligned}$$

By cutoff continuity, we get:

$$\begin{aligned} \lim_{x \rightarrow c^+} f(X) &= \lim_{x \rightarrow c^-} f(X) \\ \text{and } \mathbb{E}[\epsilon_i|X = x] &= \lim_{x \rightarrow c^-} \mathbb{E}[\epsilon_i|X = x] \end{aligned}$$

Thus, by definition of the Sharp RD LATE estimator we have:

$$\begin{aligned}\tau_{RD} \equiv y^+ - y^- &= \tau_{LATE} + \left(\lim_{x \rightarrow c^+} f(X) + \lim_{x \rightarrow c^+} \mathbb{E}[\epsilon_i | X = x] \right) - \left(\lim_{x \rightarrow c^-} f(X) + \lim_{x \rightarrow c^-} \mathbb{E}[\epsilon_i | X = x] \right) \\ &= \tau_{LATE}\end{aligned}$$

□

4.2.3 Estimation

The most common approach to estimate the treatment effect in a Sharp Regression Discontinuity Design is to run a local non-parametric regression in the neighborhood of the cutoff. It is the standard in modern RD literature to use a local linear regression to estimate the treatment effect, however, one can run a local polynomial regression as long as the degree is small, as high-order polynomial approximations give a good estimate of a function overall but perform poorly at boundary points (Runge's phenomenon). The Regression equation is written as follows:

$$Y_i = \beta_0 + \beta_1(X_i - c) + \tau_i W_i + \beta_2(X_i - c)W_i + \varepsilon_i \quad (6)$$

Where, as stated before, $W_i = \mathbb{1}_{\{X_i \geq c\}}$.

This neighborhood is defined by the bandwidth h and this regression is run excusively on the observations that respect the following constraint: $h - c \leq X_i \leq h + c$.

4.3 Difference-in-Discontinuities

The Difference-in-Discontinuities, more commonly known as Diff-in-Disc, as the name indicates, identifies the treatment similarly to a combination of Diff-in-Diff and Regression Discontinuity designs. This section goes over the identification framework as formalized in Grembi et al. (2016), extending the assumptions for this design. First, assume that all expected potential outcomes are continuous at the threshold. The Diff-in-Disc design replaces the parallel trends assumption with

a more realistic premise for this setting: local parallel trends. The common variation over time that both the control and treatment groups would experience in the absence of treatment, is solely focused at the threshold.

4.3.1 Setup

We model the potential outcomes as modeled in Butts (2023):

$$Y_{it} = f_t(X_i) + \tau_{it} \cdot W_{it} + \Gamma_i(X_i) \mathbb{1}_{\{X_i \geq c\}} + \varepsilon_{it} \quad (7)$$

- $f_t(X_i)$ summarizes location-specific characteristics that affect outcomes. They can vary across periods.
- $W_{it} = \mathbb{1}_{\{X_i \geq c\}} \mathbb{1}_{\{t=1\}}$ is the treatment variable.
- τ_{it} represents the treatment effect
- $\Gamma_i(X_i)$ denotes time-invariant, cutoff-discontinuous determinants of Y_{it} . It accounts for possible confounding policies occurring at $x = c$ as in Grembi et al. (2016).
- ε_{it} denotes other potentially unobserved individual-specific factors on the outcome at the period t .

4.3.2 Identification

Grembi et al. (2016) first introduced the particular identification assumptions for the Diff-in-Disc design. Firstly, the continuity of all potential outcomes at $x = c$ in both periods. Secondly, it introduces the assumption of a constant effect of the confounding policy over time in the absence of the treatment of interest (similar to parallel trends). It finally states that the effect of the treatment at the threshold does not depend on the confounding policy, meaning that the Diff-in-Disc estimator identifies the (local) causal effect of the treatment of interest.

Butts (2023) extends the Diff-in-Disc assumptions from Grembi et al. (2016) to a geographic setting. It states that the key for identifying the treatment effect is mainly

the hypothesis of continuity applied to $f_t(X_i)$ as well as to other possibly unobserved individual characteristics, $\mathbb{E}[\varepsilon_{it}|X_i = x]$, at the cutoff. These assumptions require that no other policies are implemented simultaneously or that effects from previously instated policies have already fully developed. Continuity of individual characteristics requires that no additional sorting can occur between 0 and 1.

I present an identification framework that relaxes continuity assumptions while maintaining local parallel trends in a single idea.

Assumption 6 (Derivative Equivalence). *The derivatives of the regression functions are equal at a neighborhood of c . Let $g_t(x) = \mathbb{E}[Y_{it}|X_i = x]$.*

$$g'_t(c^+) = g'_t(c^-) \quad (8)$$

Remark. Assumption 6 implies that:

- i. $\mathbb{E}(Y_{1it}|X_i = x)$ and $\mathbb{E}(Y_{0it}|X_i = x)$ are parallel at the threshold, validating the Local Parallel Trends Assumption.
- ii. The regression functions $\mathbb{E}(Y_{1it}|X_i = x)$ and $\mathbb{E}(Y_{0it}|X_i = x)$ are likely continuous at $x = c$, however, we cannot definitively assume continuity from this condition.
- iii. $g'_1(c^+) - g'_0(c^+) = g'_1(c^-) - g'_0(c^-) \Leftrightarrow g'_1(c^+) - g'_1(c^-) = g'_0(c^+) - g'_0(c^-)$.

This key identification assumption is very practical in the sense that the functional space where the two regression functions are continuous at the cutoff is smaller than one where the derivatives of the two regression functions are equal at the cutoff. This said it is more realistic to come across a scenario where the derivatives of the regression functions are equal at the cutoff.

Theorem 4.3 (Diff-in-Disc Identification). *Under Derivative Equivalence (A6), the Local Average Treatment Effect can be identified by the Diff-in-Disc estimator:*

$$\tau_{DiRD} \equiv (y_1 - y_0)^+ - (y_1 - y_0)^- \quad (9)$$

Where $y_t^\pm \equiv \lim_{x \rightarrow c^\pm} \mathbb{E}[Y_{it}|X_i = x]$

Proof. As previously stated, $y_t^\pm \equiv \lim_{x \rightarrow c^\pm} \mathbb{E}[Y_{it}|X_i = x]$

So all potential outcomes can be written as follows:

For units above the cutoff:

$$\begin{aligned} y_1^+ &= \lim_{x \rightarrow c^+} \mathbb{E}[Y_{i1}|X_i = x] \\ &= \lim_{x \rightarrow c^+} \mathbb{E}[f_1(X_i) + \Gamma_i(X_i) + \tau_{i1} + \varepsilon_{i1}|X_i = x] \\ &= \lim_{x \rightarrow c^+} \mathbb{E}[f_1(X_i)|X_i = x] + \mathbb{E}[\Gamma_i(X_i)|X_i = x] + \mathbb{E}[\tau_{i1}|X_i = x] + \mathbb{E}[\varepsilon_{i1}|X_i = x] \\ &\quad \text{by the linearity of expectation} \\ &= \lim_{x \rightarrow c^+} f_1(x) + \Gamma_i(x) + \mathbb{E}[\tau_{i1}|X_i = x] + \mathbb{E}[\varepsilon_{i1}|X_i = x] \end{aligned}$$

and

$$\begin{aligned} y_0^+ &= \lim_{x \rightarrow c^+} \mathbb{E}[Y_{i0}|X_i = x] \\ &= \lim_{x \rightarrow c^+} f_0(x) + \Gamma_i(x) + \mathbb{E}[\varepsilon_{i0}|X_i = x] \end{aligned}$$

For units below the cutoff:

$$\begin{aligned} y_1^- &= \lim_{x \rightarrow c^-} \mathbb{E}[Y_{i1}|X_i = x] \\ &= \lim_{x \rightarrow c^-} f_1(x) + \mathbb{E}[\varepsilon_{i1}|X_i = x] \end{aligned}$$

and

$$\begin{aligned} y_0^- &= \lim_{x \rightarrow c^-} \mathbb{E}[Y_{i0}|X_i = x] \\ &= \lim_{x \rightarrow c^-} f_0(x) + \mathbb{E}[\varepsilon_{i0}|X_i = x] \end{aligned}$$

Thus, the first differences yield:

$$\begin{aligned} (y_1 - y_0)^+ &= \lim_{x \rightarrow c^+} (f_1(x) + \Gamma_i(x) + \mathbb{E}[\tau_{i1}|X_i = x] + \mathbb{E}[\varepsilon_{i1}|X_i = x]) \\ &\quad - \lim_{x \rightarrow c^+} (f_0(x) + \Gamma_i(x) + \mathbb{E}[\varepsilon_{i0}|X_i = x]) \\ &= \lim_{x \rightarrow c^+} (f_1(x) - f_0(x) + \mathbb{E}[\tau_{i1}|X_i = x] + \mathbb{E}[\varepsilon_{i1}|X_i = x] - \mathbb{E}[\varepsilon_{i0}|X_i = x]) \end{aligned}$$

for units above the cutoff

and

$$(y_1 - y_0)^- = \lim_{x \rightarrow c^-} (f_1(x) - f_0(x) + \mathbb{E}[\varepsilon_{i1}|X_i = x] - \mathbb{E}[\varepsilon_{i0}|X_i = x]) \text{ for units below the cutoff}$$

Finally, the second difference yields:

$$(y_1 - y_0)^+ - (y_1 - y_0)^- = \tau_{LATE} + \lim_{x \rightarrow c^+} (f_1(x) - f_0(x) + \mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x]) \\ - \lim_{x \rightarrow c^-} (f_1(x) - f_0(x) + \mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x])$$

By equality of the derivatives of the regression functions on both sides of the cutoff, we can deduce that the components of the regression functions at a neighborhood of the cutoff are parallel to one another.

We then have:

$$\lim_{x \rightarrow c^+} (f_1(x) - f_0(x)) = \lim_{x \rightarrow c^-} (f_1(x) - f_0(x))$$

and:

$$\lim_{x \rightarrow c^+} (\mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x]) = \lim_{x \rightarrow c^-} (\mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x])$$

Thereby:

$$(y_1 - y_0)^+ - (y_1 - y_0)^- = \tau_{LATE} + \lim_{x \rightarrow c^+} (f_1(x) - f_0(x) + \mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x]) \\ - \lim_{x \rightarrow c^-} (f_1(x) - f_0(x) + \mathbb{E} [\varepsilon_{i1} - \varepsilon_{i0} | X_i = x]) \\ = \tau_{LATE}$$

□

5 Diff-in-Disc Estimation

5.1 Local Linear Regression

Similarly to RDD estimation of the treatment effect (Cattaneo et al., 2020), a non-parametric local linear regression at a neighborhood of the threshold is used. The neighborhood is defined by a bandwidth length h at both sides of the cutoff to limit the observations used for the regression such that only units in the interval $[c - h, c + h]$ are taken into account. The non-parametric nature of this estimation method requires the implementation of a Kernel function to add weights to the observations in the regression.

We use the equation from Grembi et al. (2016) for the Local Linear regression. That is:

$$Y_{it} = \alpha_0 + \alpha_1 X_i + \mathbb{1}_{\{X_i \geq c\}} (\beta_0 + \beta_1 X_i) + \mathbb{1}_{\{t=1\}} (\gamma_0 + \gamma_1 X_i) + W_{it} (\tau + \kappa_0 X_i) \quad (10)$$

With the treatment variable being an interaction between the treatment period indicator and the treatment group indicator $W_{it} = \mathbb{1}_{\{X_i \geq c\}} \mathbb{1}_{\{t=1\}}$.

5.2 Optimal Bandwidth

5.2.1 Plug-in Estimator (Imbens & Kalyanaraman, 2012)

For estimation of the treatment effect, a bandwidth with optimal properties is ideal to run a non-parametric Local Linear Regression at a neighborhood of the cutoff. As previously mentioned, the bandwidth proposed by Imbens and Kalyanaraman (2012) mainly has asymptotically optimal properties by constructing a bandwidth estimator that minimizes the Asymptotic Mean Squared Error. This bandwidth estimator is based on an asymptotic expansion of the MSE defined as a function of the bandwidth h and is computed in three stages:

1. Estimation of the density function of X evaluated at the cutoff, $\hat{f}(c)$, and the conditional variances above and below the cutoff, $\hat{\sigma}_+^2(c)$ and $\hat{\sigma}_-^2(c)$, respectively. These are estimated in a neighborhood determined by the Silverman Rule of Thumb Pilot Bandwidth, h_1 .

2. Estimation of the second-order derivatives of the regression function above and below the cutoff. Specifically, this involves estimating the second-order derivatives of the conditional expectation function: $\hat{m}_+^{(2)}(c)$ and $\hat{m}_-^{(2)}(c)$. These estimates are obtained via a polynomial regression within the neighborhood defined by the step 1 estimator.
3. Estimation of the regularization terms \hat{r}_+ and \hat{r}_- . These terms are computed to mitigate concerns that errors in estimating the curvature could lead to overly large and poorly performing bandwidths. This also prevents scenarios where the optimal bandwidth estimator becomes unfeasible due to a null denominator (as a direct consequence of $\hat{m}_+^{(2)}(c) = \hat{m}_-^{(2)}(c)$).

This optimal bandwidth estimator has several desirable properties:

- Consistency, regardless of the curvature at the cutoff.
- Convergence.
- Li-optimality.

5.2.2 Robust Bias-Corrected Estimator (Calonico et al., 2014)

There is also the optimal bandwidth proposed by Calonico et al. (2014), a bandwidth created for forming inference-optimal confidence intervals.

5.2.3 Adaptive Stochastic Gradient Descent

I propose a Stochastic Gradient Descent (SGD) Algorithm to obtain the MSE-optimal bandwidth, all the while using a quantile partition in the dataset to help the algorithm converge in the lower bandwidth values. In addition, the MSE is weighed by penalty terms that punish very small bandwidths to avoid ending up with very small samples that yield very bad standard deviations due to the lack of observations. This approach is practical in the sense that an SGD requires a lot less computational power than a regular Gradient Descent algorithm. This SGD algorithm uses an AdaGrad learning rate to dynamically decay the learning rate according to the data.

The objective function is defined as

$$J(h) = p(h) \cdot \text{MSE}(h)$$

Where $p(h)$ is the penalty term associated with the value of the bandwidth h . The penalty term is a function of the bandwidth in the sense that it is $\frac{N}{n_h}$ or a non-linear transformation of this term, where n_h is the number of observations such that $X_i \leq h$.

The type of the penalty term can be chosen from the following non-linear transformations:

- “default”: $\frac{N}{n_h}$
- logarithmic - “log”: $\log\left(\frac{N}{n_h}\right)$
- square root - “sqrt”: $\sqrt{\frac{N}{n_h}}$
- cube root - “cbrt”: $\left(\frac{N}{n_h}\right)^{1/3}$
- exponential - “exp”: $\exp\left\{\alpha \cdot \frac{N}{n_h}\right\} - 1$, with $\alpha = 0.01$
- sigmoid - “sig”: $(1 + \exp\{-\alpha \cdot \frac{N}{n_h} - \beta\})^{-1}$, with $\alpha = 0.05$, $\beta = 10$

In this context, there is no analytical form of the MSE function or its gradient. Therefore, the penalized MSE Gradient for a bandwidth ‘ h ’ is computed using a variation of the Finite Difference Method for Gradient calculation:

$$\nabla J(h_j) = \frac{J(h_j) - J(h_{j+})}{h_j - h_{j+}}$$

Where h_{j+} is the closest bandwidth value to h_j , such that $h_{j+} > h_j$

I use an AdaGrad learning rate for the SGD algorithm. Duchi et al., 2011 introduced this method, which often improves convergence performance over standard stochastic gradient descent in settings where data is quite sparse and is very commonly used in Machine Learning settings and has been proven to be effective for non-convex optimization problems (Défossez et al., 2020). The AdaGrad Learning Rate at the j^{th} iteration is given by:

$$\eta_{j+1} = \log(N_q) \frac{\eta_j}{\sqrt{\sum_{i \leq j} (\nabla_i J(h_i))^2}}$$

Where N_q is the size of the subset within the current h quantile interval.

Explicitly, the AdaGrad learning rate is the ratio between the previous iteration's learning rate and the Euclidean norm of the gradients from all previous iterations. I added the term $\log(N_q)$ to make up for the penalty term in the gradient as well as allowing the learning rate to be larger in bigger (more stable) parts of the function.

The SGD updates are computed as follows:

$$h_{j+1} = h_j - \eta_j \cdot \nabla_{h_j} J(h_j)$$

The algorithm stops if:

- The maximum number of iterations is reached (default = 10 but it can be modified):

$$j = \text{max_iter}$$

- The bandwidth value remains unchanged for two or more iterations (probably indicating a local minimum):

$$h_{j+1} - h_j \approx 0$$

- the difference in penalized MSE between two iterations is lower than $\frac{1}{\log(N)}$:

$$\text{MSE}_{j+1} - \text{MSE}_j < \frac{1}{\log(N)}$$

Algorithm 1 Adaptive SGD Pseudo-code

Input: $y, x, c, t, t_0, ID, kernel = "triangular", penalty_type = "default", max_iter = 10, max_epochs = 500$

Output: $optimal_h, optimal_mse$

Initialization:

Define q quantile intervals based on sample size N

for each q **do**

for i from 1 up to max_epochs **do**

Initialize bandwidth by choosing a random h_0 and compute $\nabla_{h_0} J(h_0)$

for j from 1 to max_iter **do**

Compute the learning rate using AdaGrad updates

Update bandwidth: $h_{j+1} \leftarrow h_j - \eta_j \cdot \nabla_{h_j} J(h_j)$

Calculate $J(h_j)$ and $\nabla_{h_j} J(h_j)$

if stopping criteria met (e.g., minimal MSE or change unchanged h_j)

then

Break iteration

end if

end for

Save h_i and $J(h_i)$

end for

end for

Choose $h^* = \min_i J(h_i)$

Return $h^*, J(h^*)$

In the next section, I present the results for comparison between the optimal bandwidth estimator from Imbens and Kalyanaraman (2012), the bias-corrected optimal bandwidth estimator from Calonico et al. (2014), and the optimal bandwidth obtained by the SGD algorithm, with triangular kernel weights, as recommended in Imbens and Kalyanaraman (2012), and uniform kernel weights. These three estimation methods will be implemented in a simulated database in order to

test their performance on the simplest of data. These methods will be later applied to real-world data.

6 Simulation

6.1 Data

To illustrate the effectiveness of the Diff-in-Disc method, I simulated a database in the context of a geographic Diff-in-Disc where the two studied regions are Paris and Issy-les-Moulineaux. The simulation consists of a panel composed of $N = 10000$ units, representing different properties around the Paris / Issy-les-Moulineaux area, observed through $t = 6$ years, leaving us with a total of $6N = 60000$ observations. Butts (2023) extends the assumptions made in Keele and Titiunik (2015) to the Diff-in-Disc framework. In this particular application, I assume the use of a scalar function, in this case, the Euclidian distance of a unit i to the border. The observations are distributed along an $(X; Y) \in \mathbb{R}^2$ axis such that their Euclidian distance to 0 represents the distance to the border. Properties with a negative `Y_axis` value are located in the Issy-les-Moulineaux area and the ones with a positive `Y_axis` value are located in Paris. The running variable is the distance to the cutoff c , representing the border between two geographic regions.

The simulated `Price_m2it` as the variable of interest representing the price of the squared meter of the property i at the year t . `Price_m2it` is determined by 4 factors and an intercept. Treatment is assigned to observations in Paris (`Y_axis` ≥ 0) from the fourth year onwards. The four determinants of the price of the squared meter are varying in characteristics:

1. Common price determinant in both areas: an indicator if the property has a balcony or not.
2. Border-continuous price determinant: the Cartesian distance to the Eiffel Tower, assuming it to be at the coordinates (0;3400).
3. Border-discontinuous determinant in the price: access to the metro (I assume that Paris has metro access whilst Issy does not).
4. Price determinant following a Local Parallel Trend: real estate speculation.
We assume that there's a similar growth in real estate prices in the neighbor-

hood around the border and that it's a very different growth than closer to downtown Paris or further South in Issy.

Thus, the equation determining the price of the squared meter is:

$$\text{Price_m2}_{it} = \beta_0 + \beta_1 * \text{Balcony_indc}_i + \beta_2 * \text{Distance_to_EiffelTower}_i \\ + \beta_3 * \text{Metro_access}_i + \beta_4 * \text{Local_Trend}_{it} + \tau * \text{Treatment}_{it} + \varepsilon_{it}$$

With:

- $\beta_0 = 10000$
- $\beta_1 = 500$
- $\beta_2 = -1.7$
- $\beta_3 = 200$
- $\beta_4 = 50$
- $\tau = 350$ (Treatment effect)

The nature of these determinant factors in the variable of interest makes it such that using Diff-in-Diff or RDD methods for estimation of the treatment effect is inherently biased, as well as a regression over the entire sample using Equation 10 (Grembi et al., 2016). In theory, our “real estate speculation” variable violates the Parallel Trends assumption for Diff-in-Diff identification, and the “access to the metro” variable acts as a confounding factor violating the continuity assumption necessary for RD identification. Thus, by design, our simulated data requires the use of the Difference-in-Discontinuities method to correctly estimate the treatment effect.

In the regression, an asymmetric noise that differs between periods is introduced such that the noise level in the post-treatment period is twice as strong as in pre-treatment.

6.2 Results

I begin by empirically showing that both Diff-in-Diff and RDD methods are biased in estimating the treatment effect, and then I compare the results of three optimal bandwidth estimation approaches:

1. The optimal bandwidth estimator presented in Imbens and Kalyanaraman (2012)
2. The average of two optimal bandwidth estimators used in Grembi et al. (2016)
3. The minimal-MSE bandwidth found by our Adaptive SGD algorithm

I also present the results for two different kernels: uniform and triangular.

6.2.1 Biased Results for RD and DiD

In theory, the identification of the treatment effect for both Diff-in-Diff and Regression Discontinuity is biased. First, the Diff-in-Diff design cannot be applied to this framework because the overall trends for both groups are not parallel, breaking the key identifying assumption. As for RDD, the presence of a border-discontinuous confounder violates the identifying assumption. Consequently, both of these methods produce biased results when estimating the treatment effect.

Lastly, when doing a Linear Regression using Equation 10 on the overall sample, the fact that high-order polynomial approximations give a good estimate of a function overall, but perform poorly at boundary points becomes evident (as specified in Cattaneo et al., 2020) when the goal is to estimate the regression functions at a boundary point. That is to say, when the point estimator is affected by observations far from the boundary, it is not so pertinent to use global polynomial approximations.

Remark. Notice how the bias for the estimation of the treatment effect in table 1 is equal to the coefficient associated with the border discontinuous determinant in the price.

	Diff-in-Diff	RDD	Full sample Diff-in-Disc
Coefficient	1049.24***	553.51***	918.51***
Bias	699.24	203.51	568.51

Table 1: Estimation of the Treatment Effect and Respective Bias

6.2.2 Diff-in-Disc results for different Bandwidth Estimators

The first estimator is the plug-in optimal bandwidth estimator from Imbens and Kalyanaraman (2012) for Regression Discontinuity Designs. This estimator allows the construction of the same bandwidth in both sides of the regression.

Grembi et al. (2016) estimates the treatment effect by running a local linear regression in a neighborhood of the cutoff. The sample size to run the local linear regression is obtained by averaging the two bandwidths obtained with the `rdrobust` package from Calonico et al. (2015): one bandwidth for the pre-treatment period and another one for post-treatment.

Lastly, the Adaptive SGD finds the bandwidth value that minimizes the Mean Squared Error associated with Equation 10.

Kernel	Imbens Plug-in Estimator		AVG of 2 Bandwidths		Adaptive SGD	
	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform
Bandwidth	81.26	127.65	184.38	148.83	114.80	114.11
Treatment	347.51***	356.09***	353.70***	355.47***	349.74***	352.30***

Table 2: Estimated Treatment Effect for Different Bandwidth Methods and Kernels

Firstly, the simulation shows that the Diff-in-Disc method is, in practice, very effective at correctly identifying the treatment effect while avoiding biases that originate from a discontinuity at the cutoff and does not require the strictness involved with the parallel trends assumption to be effective. In addition, the Adaptive SGD Algorithm was the most precise when estimating the treatment effect.

7 Application - Estimating the Effect of Rent Control in Paris on Real Estate Prices

7.1 Context

Paris has been struggling with an ongoing housing crisis, where skyrocketing real estate prices have made it increasingly difficult for residents to afford homes or tenants to afford rent in a very limited space. This issue stems from high population density, limited land for development, and a strong influx of global capital into Paris's real estate market. These factors have led to persistent upward pressure on rental and sale prices, causing problems for residents across the socioeconomic spectrum. In Paris, where demand for rental properties far exceeds supply, the introduction of rent control (*encadrement des loyers*) represents a crucial policy intervention aimed at curbing the rapid increase in rental costs and ensuring more equitable access to housing. However, the economic impacts of such policies, particularly on property prices, remain contentious.

The implementation of rent control policies has long been a topic of debate among economists and policymakers, particularly in densely populated urban areas, where housing affordability is a significant concern. This study is motivated by the need to evaluate the effect of rent control on the price per square meter in the Paris real estate market. Understanding these effects is essential for evaluating the overall effectiveness of rent control, as a tool for addressing housing affordability, and for guiding future policy decisions in Paris and other similar urban settings.

The effects of rent control on housing markets have been widely studied, with findings that vary depending on the context and specific design of the policy. Rent control is often credited with providing immediate relief to tenants by capping rent increases, thus preventing displacement in high-demand areas. However, numerous studies have highlighted potential downsides, including reduced incentives for landlords to maintain and invest in their properties and a decrease in the overall supply of rental housing in an over-saturated demand.

For instance, Gyourko and Linneman (1989) conducted an influential study on the impacts of rent control in New York City, examining both the equity and efficiency aspects of the policy. They found that rent control, while often politically popular due to its tenant protections, can have adverse effects on property values. In particular, their study highlighted that controlled rents reduce the attractiveness of rental properties to investors, thereby depressing property values over time. This dynamic is especially pertinent to the Paris real estate market, where high demand and tight supply mean that any reduction in the incentives for property investment could result in lower transaction prices.

On the other hand, some studies suggest that the design of rent control policies matters greatly. For example, Autor et al. (2014) argue that carefully crafted rent control measures (e.g. allowing for some market adjustments rather than imposing rigid caps on rent) can protect tenants without significantly disrupting the overall housing market.

Turner and Malpezzi (2003) provide a comprehensive review of the empirical evidence on rent control policies, highlighting both the short-term benefits and long-term inefficiencies. While rent control can provide immediate relief to tenants by capping rent increases and protecting them from sudden price hikes while making the housing market inefficient (as mentioned in Gyourko and Linneman, 1989). In contrast, more flexible policies can mitigate some of these negative outcomes. This framework is highly relevant to the Parisian market, where rent control may be affecting transaction prices in distinct ways across different neighborhoods, depending on local market conditions.

Given this mixed evidence, the impact of rent control on prices, particularly in a complex and over-saturated urban market like Paris, remains an open question. This study contributes to the literature by providing empirical evidence on how the *encadrement des loyers* policy in Paris has influenced the price per square meter of real estate as well as being the first study to use a geographical Diff-in-Disc design to evaluate territorial public policies.

According to the official website of the French administration, rent control is a device that limits the rise in rent when signing or renewing a lease. It applies

to municipalities where the number of available rental properties is far below the number of people who want to become tenants and make these properties their primary residence. In Paris, the rent for a property with a lease signed or renewed since July 2019 is regulated by lower and upper reference rents. These amounts vary depending on the type of rental (unfurnished or furnished), the number of rooms, and the period of the property's construction, as Paris was the first of several regions where this rent control device was put into place. In June of 2021, the same policy was applied in Plaine Commune which consists of the northern *communes* (or municipalities) from the Seine-Saint-Denis department. Later in December of 2021 the Eastern municipalities of Seine-Saint-Denis known as Est-Ensemble adopted this same policy. Hence, the study focuses on data before June 2021.

I use the Diff-in-Disc framework on Paris and the surrounding municipalities to know to what extent the rent control policy in Paris affected real estate prices.³

7.2 Data

I apply the Diff-in-Disc method on the Demandes des Valeurs Foncières (DVF) géolocalisées database which contains all Real Estate transfers that have been performed in a year (since 2019). For each observation, this database contains the date of the transfer, the land value, the surface, and the address, including the postal code, the name and code of the department, and the *commune*. It also indicates the type of transfer (sale, exchange, expropriation, sale in the future state of completion, adjudication, sale of land for building) as well as the type of building (apartment, house, industrial or commercial complex, dependency). All of these transfers also contain the longitude and latitude of the building which will allow the calculation of distances from the properties to the border for the running variable.

³The surrounding municipalities that share a border with Paris: Aubervilliers, Bagnolet, Boulogne-Billancourt, Charenton-le-Pont, Clichy, Fontenay-sous-Bois, Gentilly, Issy-les-Moulineaux, Ivry-sur-Seine, Joinville-le-Pont, Le Kremlin-Bicêtre, Levallois-Perret, Les Lilas, Malakoff, Montreuil, Neuilly-sur-Seine, Nogent-sur-Marne, Pantin, Puteaux, Saint-Cloud, Saint-Denis, Saint-Mandé, Saint-Maurice, Saint-Ouen, Suresnes, Vanves, Vincennes.

The administrative borders are obtained *Communes d'Île de France* shapefile that includes the territorial division of all *communes* in Île de France. With the help of the **sf** or simple features package for R, we traced the border limits of Paris.

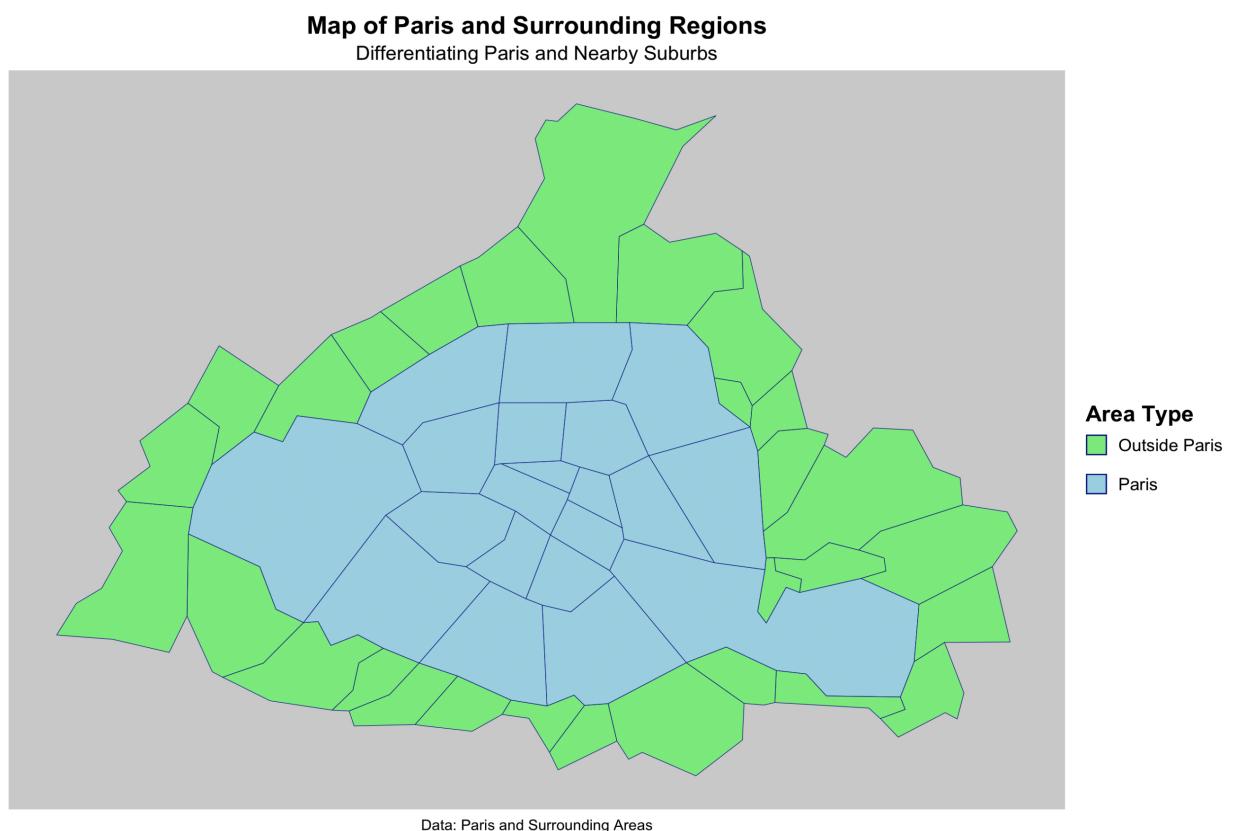


Figure 1: Studied Area color-coded by Treatment/Control group

Since Earth is a sphere, a simple application of the Euclidean distance with the points defined by latitude and longitude would not be sufficient. The `simple features` package uses the geodetic distance to match the format from the *Communes d'Île de France* shapefile. To do this, first, the borders need to be converted to a line string: if not, the observations inside the city of Paris will all be considered to have a distance to the border equal to zero as they are inside of the polygon.

Thus, I add to the DVF database the “distance to the border” variable for each of the observations. For Diff-in-Disc analysis, the values from this variable are positive in the treatment region and negative in the control region.⁴

For the area of interest and the years 2019, 2020, and 2021 (up until June), aggregating a total of 283,867 observations before data cleaning. For the analysis, all units under study must include both longitude and latitude coordinates. That leaves out all that are missing one of these values amounting to 0.92% of observations. Additionally, there are instances where a single transaction involving multiple surfaces is registered as multiple entries. To address this, I regroup these “multiple” transactions into a single observation aggregating the surface values by transaction ID and averaging the distance to the border of the surfaces. This process reduces the size of the database by 18%.

I winsorize the top 3% and bottom 3% of the observations. It is of most pertinence in this context because it allows for the reduction of the influence of outliers while preserving the integrity of the data. This is a technique used to deal with outliers in the data by setting, in this case, the values above the 97th percentile and below the 3rd percentile are replaced by the values at these cutoffs. This approach is particularly useful when minimizing the impact of outliers without completely discarding data, as it preserves the overall structure of the dataset while reducing the influence of extreme values. As this study revolves around estimating the Local Average Treatment Effect (LATE), it is crucial to manage extreme values in a way that minimizes their impact on the overall analysis. Replacing the extreme values with more moderate ones ensures that the analysis remains centered around the average effects, leading to more robust and reliable estimates.

⁴Detailed in equation 11

The missing values (NAs) in the price per squared meter variable are replaced by using the K-Nearest Neighbors (KNN) method with $k = 5$. The KNN algorithm is a non-parametric method where, for each missing value, the algorithm identifies the five nearest neighbors in the dataset based on the available features, and the missing value is then attributed based on the average of these neighbors. This is particularly effective in capturing the local structure of the data, ensuring that the values are consistent with the observed patterns.

Finally, to simplify the data, I keep the following variables: 1) name of the commune; 2) department code; 3) transaction ID; 4) transaction date; 5) latitude; 6) longitude; 7) distance to the border; and 8) price per squared meter. This dataset contains 119,297 observations.

The observations for each year are geographically distributed as follows:

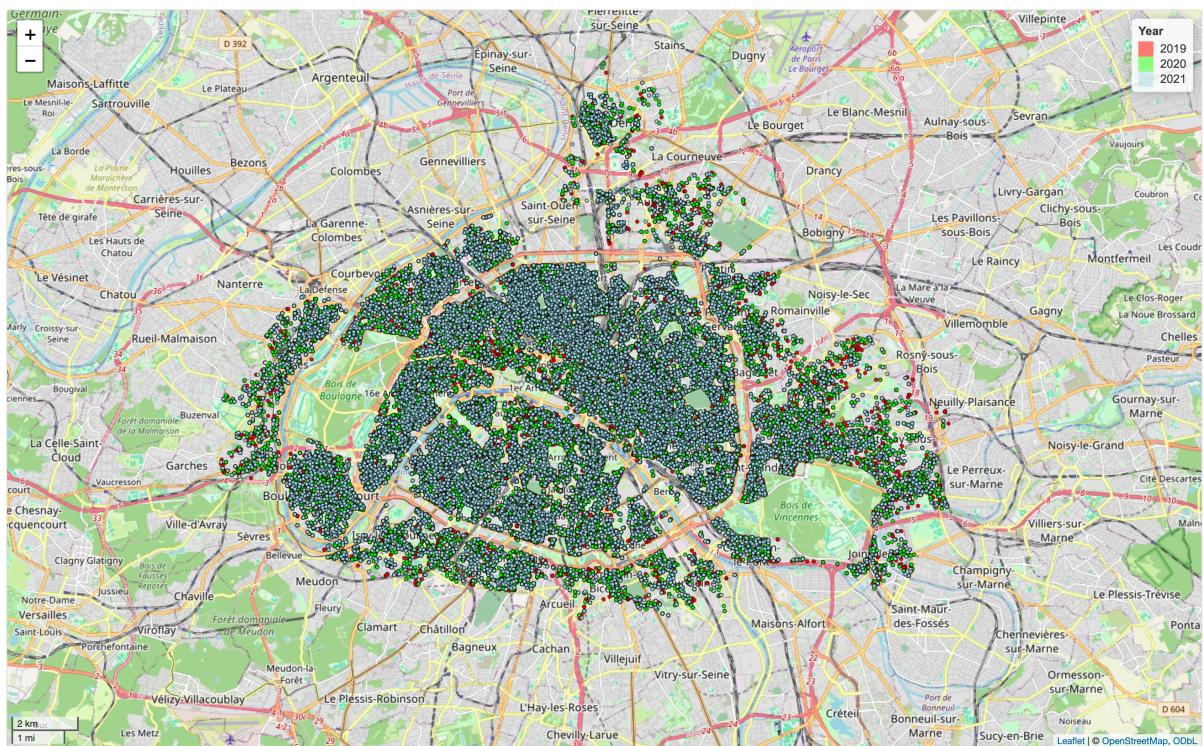


Figure 2: Geographic Distribution of Observations Grouped by Year

Let D_i represent the distance to the border from the property corresponding to the observation i . We add the running variable such that:

$$X_i = \begin{cases} D_i & \text{if } i \in \text{Paris} \\ -D_i & \text{else} \end{cases} \quad (11)$$

However, some observations find themselves outside of France ($12 * 10^7$ meters away from the border), likely due to inverted latitude and longitude coordinates. To conduct an exhaustive analysis, the sparsity of the data must be reduced. Therefore, all observations further than 100 km away from the border are eliminated thus excluding 2.13% of observations. Finally, I will be working with a data frame constituted by 116,753 observations of 8 variables.

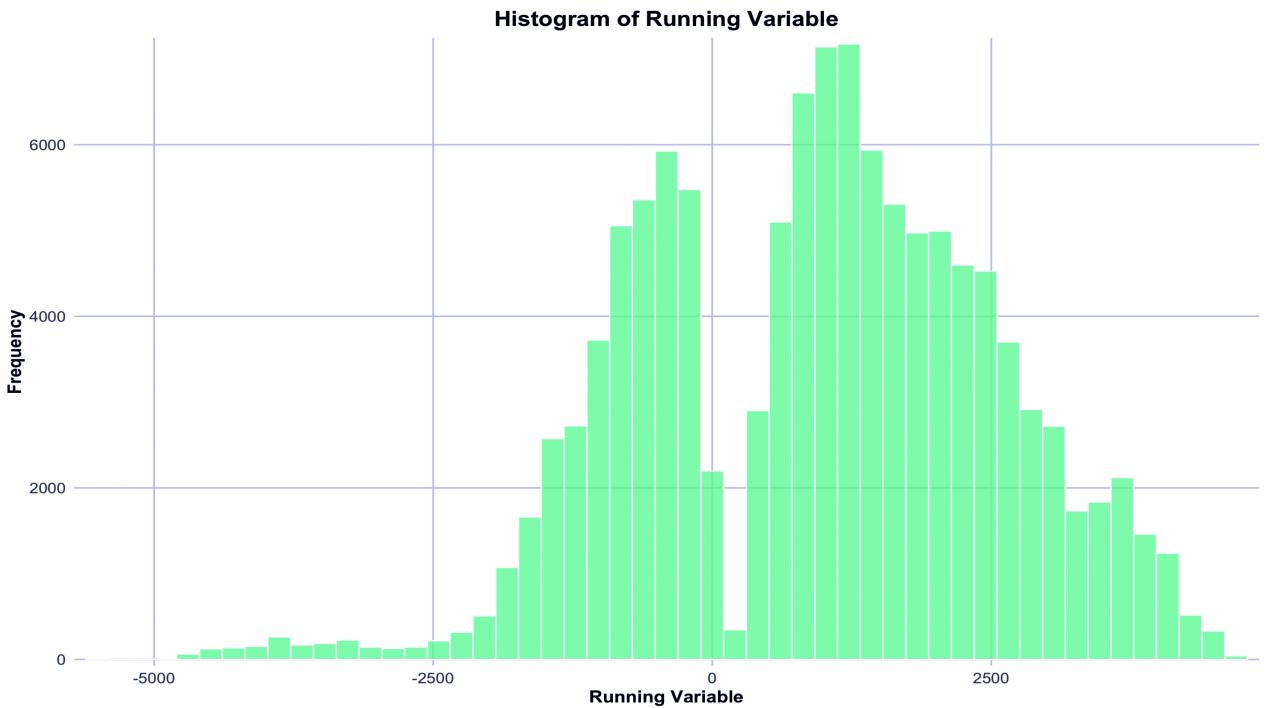


Figure 3: Histogram of the Running Variable within 100km from the border

Among these variables, I focus on a more detailed examination of the two key variables used in the Diff-in-Disc analysis: the running variable (the modified distance to the border) and the price per square meter (the outcome variable). I present general descriptive statistics for these variables, along with a more detailed analysis that explores the characteristics of specific districts and municipalities. This includes an examination of trends and the overall evolution of these areas, providing deeper insights into the dynamics at play.

Statistic	Value (€/sqm)
Count	116,753.00
Mean	9,271.34
Median	9,433.96
SD	3,495.74
Min	1,429.57
1st Quartile	7,097.56
3rd Quartile	11,425.00
Max	17,051.06
Skewness	-0.06

Table 3: Descriptive statistics for the Price per Square meter

Overall, the descriptive statistics suggest that the price per square meter in the study area is generally high, which is coherent for a very high-demand area like Paris. The slight negative skewness and relatively close mean and median values suggest that while the mean is slightly pulled upwards by extremely high prices, the distribution of prices is fairly symmetric. In addition, there is quite a broad range in prices. This is further illustrated by the histogram of the variable in figure 4. The high frequency of extreme prices is a consequence of the winsorization procedure.

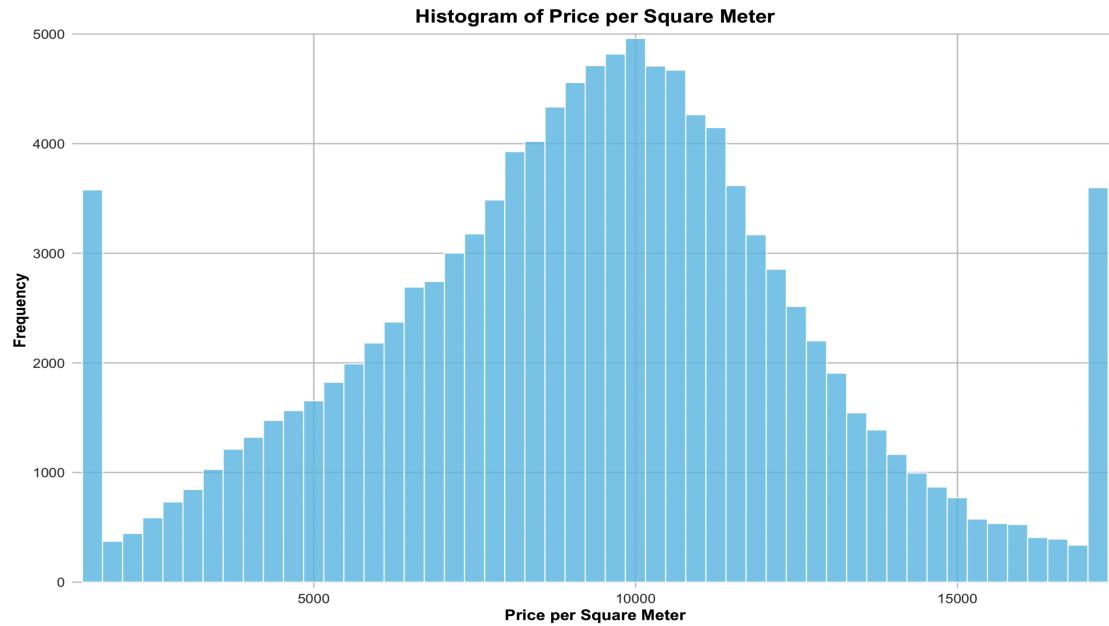


Figure 4: Histogram of the Price per Square Meter in Paris and its surrounding *Communes*

To visualize the differences in the square meter price, the distribution of the square meter in Paris can be compared with that outside Paris.

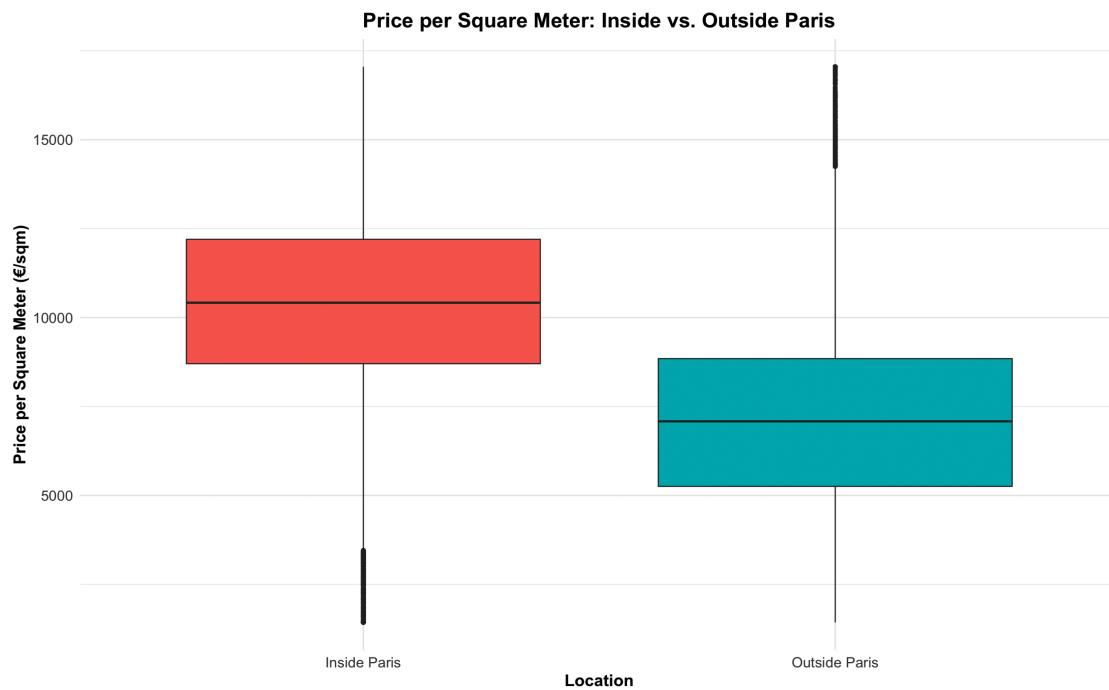


Figure 5: Comparison between Box Plots Inside and Outside of Paris

The two box plots clearly illustrate that the price per square meter is significantly higher in Paris compared to the adjacent *communes*. The median price per square meter for properties in Paris exceeds €10,000.00, while the median for properties in the *Petite Ceinture* is just below €7,500.00. Additionally, the third quartile for properties outside Paris aligns closely with the first quartile for properties within Paris, both around €9,000.00 per square meter. This stark contrast highlights the considerably higher property prices within the city of Paris compared to the surrounding areas. Furthermore, the outliers in Paris are predominantly concentrated at the lower price range, while the outliers in the surrounding *communes* are mostly found at the higher end of the price spectrum.

To continue with the exploration of the housing market in Paris and its surrounding municipalities, one can take a closer look at the general evolution through the years, illustrated by figure 6.

In 2019, before the implementation of rent control in Paris, the average and median prices per square meter inside the city were significantly higher than those outside of it. The median price inside Paris was €10,000, indicating a strong market demand. The lower prices outside Paris, with a median of €6,764.92, reflect the less competitive real estate market in the surrounding *communes*. There is no observable effect of the rent control policy in any of the two areas following its implementation in July 2019. In 2020 however, the price per square meter inside Paris increased notably, with the mean rising to €10,623.64 and the median to €10,746.21. This increase may suggest that the initial impact of rent control did not depress property prices as might have been anticipated; instead, prices continued to rise, potentially due to the high demand and limited supply within the city. The effects that the COVID-19 pandemic may have had on the evolution of the housing market in both areas. As a matter of fact, in figure 6 it can be seen that in 2020 there is an inflection point for both locations.

Outside Paris, the prices also rose, though not as sharply, with the mean and median prices increasing to €7,311.58 and €7,272.73, respectively. This modest increase could be attributed to a spillover effect, where buyers moved out of Paris and turned to surrounding *communes*. The impact of COVID-19 may have also led

to a revaluation of suburban areas, as more people sought properties outside the city during lockdowns.

By 2021, the price per square meter inside Paris showed a slight decrease in the mean to €10,588.11, but the median remained relatively stable at €10,731.81. This slight decline in the mean could indicate a stabilization of prices following the initial shock of rent control and the ongoing effects of the pandemic. However, the high median suggests that demand for property remained strong, attributable to the continued attractiveness of the city.

In the surrounding municipalities, the mean and median prices continued to increase to €7,458.70 and €7,485.28, respectively. This rise reflects the ongoing trend of people moving to suburban areas, a shift accelerated by the pandemic, as stated before, and the potential for more remote work opportunities. It could also indicate a break in trends between Paris and the outside *communes*, as Rent Control could have given the adjacent *communes* a special attractiveness for investors as the price of rent was not controlled.

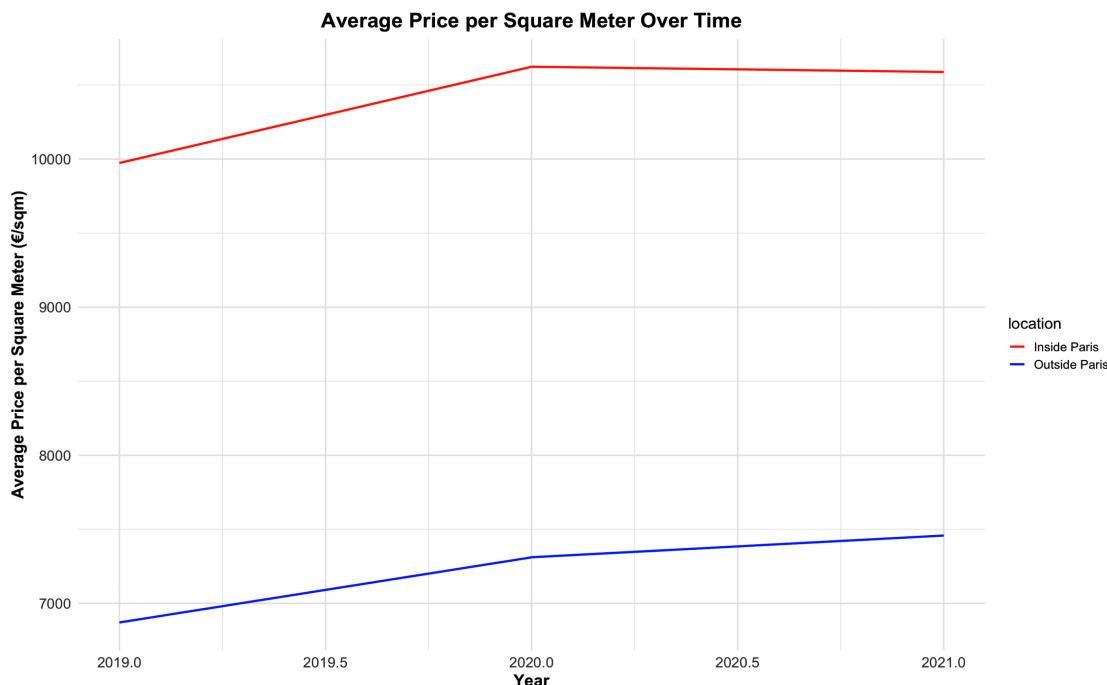


Figure 6: Average per Square Meter Price Evolution split between Regions

The next relevant thing for the study is to explore the trends in average prices per square meter over the years, it is essential to examine how these prices vary across different districts within Paris. While the overall trends provide a broad overview, they do not capture the nuanced differences between specific neighborhoods.

District	N	Median (€/sqm)	SD (€/sqm)	Q1 (€/sqm)	Q3 (€/sqm)
Paris 6th	1,975	14,339.62	4,409.57	11,697.95	17,051.06
Paris 7th	2,286	14,052.29	4,003.37	11,766.54	16,666.67
Paris 4th	1,302	12,903.23	3,975.66	10,910.99	15,000.00
Paris 1st	782	12,642.21	4,067.07	10,388.46	15,300.94
Paris 5th	2,116	12,349.30	3,663.61	10,385.13	14,167.58
Paris 3rd	1,749	12,222.22	3,956.32	10,188.68	14,074.07
Paris 8th	1,879	12,037.00	4,185.29	9,934.24	14,044.23
Paris 2nd	1,314	11,663.33	3,790.98	9,626.73	13,218.83
Paris 9th	3,000	11,344.15	3,566.98	9,482.18	12,973.78
Paris 16th	6,972	11,184.29	3,390.25	9,600.00	12,896.46
Paris 17th	6,888	10,876.24	3,148.11	9,285.71	12,292.46
Paris 11th	6,487	10,555.56	2,986.46	9,117.65	11,838.40
Paris 14th	3,866	10,331.92	2,828.02	8,995.10	11,677.14
Paris 15th	8,016	10,295.77	2,696.06	8,970.59	11,618.21
Paris 10th	4,165	10,284.34	3,130.35	8,750.00	11,617.02
Paris 12th	4,055	9,750.00	2,662.15	8,453.46	10,888.89
Paris 18th	8,551	9,666.67	2,940.81	8,030.66	11,250.00
Paris 13th	3,588	9,451.56	2,700.52	8,013.51	10,724.60
Paris 20th	4,952	9,036.04	2,372.13	7,867.98	10,192.31
Paris 19th	4,327	8,687.50	2,515.39	7,307.69	9,971.04

Table 4: Descriptive Stats for all Paris Districts

Table 4 offers a closer look at the median prices, variability, and price distribution within various Parisian districts, highlighting the distinct characteristics and market dynamics that influence values in each area. By analyzing these district-level statistics, one can better understand the spatial variations in the real estate market and how factors such as location, demand, and gentrification contribute to the broader price trends observed in the previous section. There is a clear division between districts, which can be regrouped into 6 groups.

The 6th and 7th *Arrondissements* stand out as the most expensive areas, with median prices per square meter of €14,339.62 and €14,052.29, respectively. These districts are known for their prime locations, historic architecture, and proximity to cultural landmarks like the Luxembourg Gardens and the Eiffel Tower, which contribute to their high value. The significant variability in prices suggests a wide range of property types and conditions within these areas, from luxurious apartments to smaller, perhaps less modernized units, that nonetheless retain great value due to the prestige of the neighborhood.

The 1st, 3rd, 4th, and 5th *Arrondissements* also command high prices, with median values exceeding €12,000 per square meter. The 1st district, home to the Louvre and Palais Royal, along with the historic Marais area in the 4th, reflect their desirability through high prices. The interquartile ranges in these districts are narrower compared to the top-end districts, indicating more consistent pricing across properties within these central historical areas.

The 2nd, 9th, and 16th *Arrondissements* have median prices slightly above €11,000 per square meter, placing them in the middle tier of Parisian property values. The 16th *Arrondissement*, known for its residential nature and proximity to the Bois de Boulogne, has a somewhat broader distribution of prices, as indicated by its standard deviation. The 8th is an exception in this category, having a median price of €12,037.00 and a quite high standard deviation of €4,185.29. These districts offer more affordable options compared to the very central or prestigious districts, retaining the appeal of the very desirable living areas with good access to amenities and transport.

The 10th, 11th, 14th, 15th and 17th *Arrondissements* have median prices between €10,000 and €11,000 per square meter. These neighborhoods are a blend of residential tranquility and urban convenience.

The 12th, 13th, and 18th *Arrondissements* exhibit median prices below €10,000 per square meter. The relatively higher standard deviations in these areas indicate ongoing shifts in the market, with significant price variability influenced by the specific neighborhood and proximity to transport hubs. This variability underscores the diverse and contrasted nature of these districts.

Finally, Paris' 19th and 20th *Arrondissements* are the most affordable with median prices of €8,687.50 and €9,036.04 per square meter, respectively. The lower quartiles (Q1) in these districts are also significantly below those of more central areas, indicating that there are still relatively affordable properties available, which could appeal to first-time buyers or those seeking investment opportunities in emerging neighborhoods.

The table illustrates a clear price tendency across Paris, with prices decreasing as one moves away from the central and prestigious western districts towards the eastern and northern districts. This variability must also be compared with the distances of these properties to the Paris border. As shown in table 8, it is evident that the Parisian districts closest to the border predominantly fall within the lower price tiers.

7.3 Diff-in-Disc Analysis

Using the same methods as in the simulation (Section 6), I present the results of the Diff-in-Disc model applied to the dataset, utilizing the three different bandwidth selection methods discussed earlier in this study: 1) the Imbens Plug-in Estimator; 2) the Average of Two Bandwidths; and 3) the Adaptive SGD approach. Each method is applied using both triangular and uniform kernels.

Table 5 summarizes the estimated bandwidths, treatment effects, and their associated standard deviations for each method and kernel combination.

The Plug-in Estimator produced relatively wide bandwidths, particularly when applying the uniform kernel, resulting in treatment effect estimates that are negative and vary significantly across kernel types. Due to the large variability in the estimates, this conclusion is not statistically robust.

The Average of two bandwidths obtained using the `rdrobust` package approach also yielded negative treatment effects, although their magnitude is smaller compared to the Imbens Plug-in Estimator.

Finally, the Adaptive SGD method produced the smallest bandwidths among the three approaches, leading to positive treatment effects. This would be of concern, however, because none of these effects are statistically significant. The standard deviations associated with the SGD estimates are larger than those obtained with the other methods, which may suggest increased variability due to the smaller bandwidth size.

The very large standard deviations associated with each of the estimates of the treatment effects make the effects of the policy statistically insignificant.

The first key finding of the analysis applied to the impact of the rent control policy on property prices in Paris is that there is no statistically significant evidence of a clear effect, whether positive or negative. The large standard deviations associated with the treatment effect estimates suggest that the observed variations are due to heterogeneity between the regions, yielding very high standard deviations.

	Imbens Plug-in Estimator		AVG of 2 Bandwidths		Adaptive SGD	
	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform
Bandwidth	444.43	698.15	673.56	637.13	230.96	231.69
Treatment	-510.29	-88.74	-356.45	-134.16	291.46	274.57
	(569.31)	(339.28)	(361.20)	(390.76)	(953.5)	(1117.23)

Table 5: Estimated Treatment Effects for Different Optimal Bandwidth and Kernels

Clustering

To obtain smaller standard deviations and reduce heterogeneity, the observations can be separated into clusters. This successfully reduces the internal heterogeneity and can improve the significance of the estimates. To do this, there is a very large number of methods to cluster observations.

For this study, I will employ the K-means clustering method due to its simplicity and pertinence in the geographical setting. K-means is an unsupervised machine learning algorithm that groups data points into clusters based on their similarities. The algorithm begins by selecting a set number of cluster centers, called centroids, and assigns each data point to the nearest centroid. Once all points are assigned, these centroids are recalculated as the average position of the points in each cluster.

To determine the ideal number of clusters (K) for K-means, the elbow method is often used. It involves running the algorithm for a range of K values and plotting the within-cluster sum of squares (WCSS), which is an indicator of internal heterogeneity. As K increases, the WCSS decreases, but the improvement diminishes after a certain point, forming a curve with a noticeable “elbow”. The K value at this elbow represents the optimal number of clusters, where adding more clusters no longer significantly reduces the internal Sum of Squares. This method yields $K = 2$ clusters as seen in figure 8.

It is noticeable how a pattern emerges looking at the geographical distribution of the split observations in figure 7 with prior knowledge of the distribution of prices in Paris (cf. Table 4), added to the basic statistics of both clusters (Table 6):

- The first cluster contains the less expensive regions around the Paris border, for example, the Pantin - 19e *Arrondissement* border is in this cluster.
- The second cluster contains observations in the more expensive neighborhoods adjacent to the border, illustrated by the Neuilly-Sur-Seine / Levallois-Perret - 17e *Arrondissement* area.

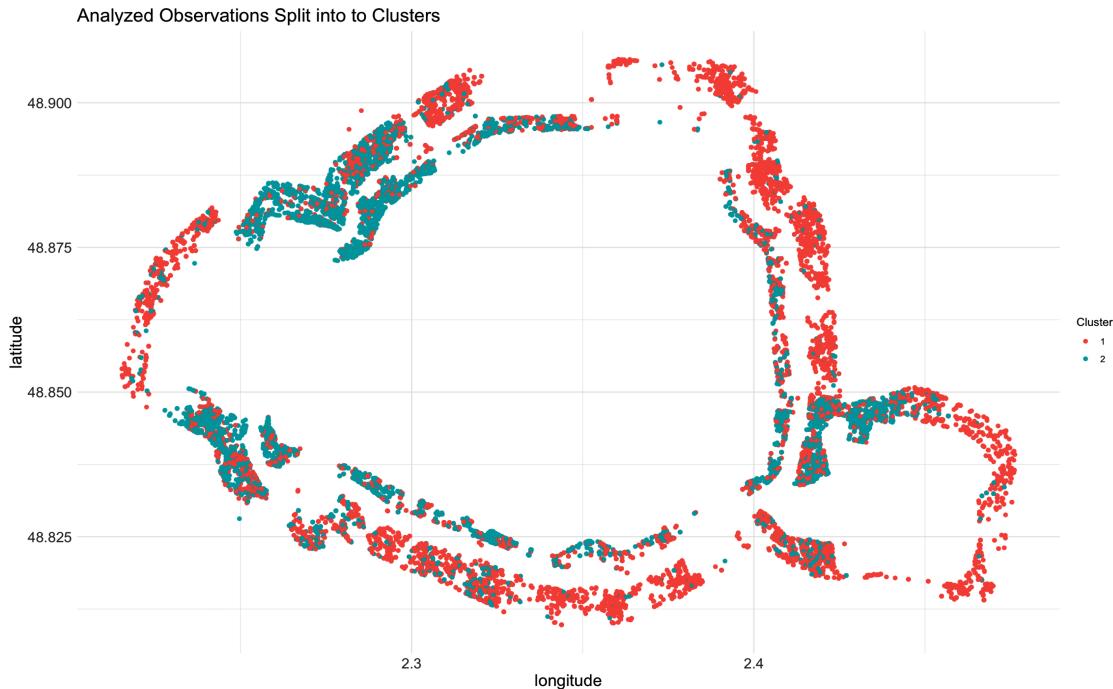


Figure 7: Geographical Distribution of Observations Split by Cluster

Cluster	Count	Mean Price (€/sqm)	Median Price (€/sqm)
1	11,875	5,830.07	6,261.36
2	13,342	10,205.46	9,761.91

Table 6: Summary Statistics by Cluster

When running a local linear regression on each cluster using equation 10, for each of the previous optimal bandwidth methods, using both the triangular and uniform kernels, the regression yields more robust estimates.

The estimates obtained by the Imbens Plug-in bandwidth are not significant for the triangular kernel but significant for the uniform kernel in both clusters. For the Average of two bandwidths from the `rdrobust` package, the estimates are robust for the second cluster using both kernels, while in the first cluster, only the triangular kernel yields robust estimates. Regarding the Adaptive SGD optimal bandwidth, the only statistically significant estimate is for the second cluster using the triangular kernel.

Generally, there is a tendency in the sign of the estimates: all in the first cluster are negative, while all in the second cluster are positive. It is also remarkable that the optimal bandwidth method yielding the more robust estimates is the Average of Two Bandwidths.

The second key finding indicated by these results is that the effect of the *Encadrement des Loyers* policy is not homogeneous. The estimates indicate that the price of the squared meter in lower-valued neighborhoods decreased, which is in sync with the findings from Gyourko and Linneman (1989), while the effect was the opposite in higher-valued neighborhoods.

	Imbens Plug-in Estimator		AVG of 2 Bandwidths		Adaptive SGD	
	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform
Cluster 1	-233.40	-370.90	-426.60*	-341.50	-132.64	-132.83
	(266.70)	(207.00)	(211.40)	(215.90)	(410.67)	(410.81)
Cluster 2	323.53	539.32**	452.65**	467.02**	466.90*	454.90
	(229.85)	(171.68)	(174.27)	(180.56)	(338.84)	(338.90)

Table 7: Estimated Treatment Effects for different Clusters

8 Conclusion

This study proves quite useful when evaluating territorial public policies, especially in heterogeneous markets such as Paris. My findings reinforce that public interventions are not magical tools that have the same effect on the entirety of the population. Indeed, the effect of territorial public policies is subject to the context of certain markets and it would be more appropriate for policymakers to focus on more concentrated communities or municipalities so as not to have unforeseen effects on certain sectors of the same city.

“...all of economics is a behavioral science. The difference is whether you think the behavior is irrational or rational.”⁵

Moreover, this study underscores the need for robust statistical tools when evaluating policy impacts. Without dividing the sample into clusters, the estimates were non-significant, leaving the application of this method inconclusive. Besides, clustering is quite important in this context because when adding positional covariates to try to mitigate the noise in the standard errors, a multicollinearity issue between a positional indicator and the running variable would arise.

The analysis revealed that the Diff-in-Disc approach, while powerful, is sensitive to the choice kernel, which significantly influences the estimated treatment effects. This study also highlighted the inherent challenges in dealing with real-world data, such as errors in data collection (inverting latitude and longitude, missing values...), outliers, and the impact of external factors. Plus, there are other clustering algorithms that could be used in this context (i.e. `max_p_tabu` algorithm from the `rgeoda` package).

Furthermore, the application results evidence that the Adaptive SGD algorithm tends to find bandwidths leading to estimates with very large standard deviations. This might need future reinforcement to try to produce robust estimates if one is to strengthen the capacities of the algorithm. There are other learning rates for non-convex optimization problems that could be built into the Adaptive SGD.

⁵Eugene Fama, *Financial Times*, August 20, 2024

While the current study has focused on a standard Diff-in-Disc model, several extensions could enhance and expand the applicability of the Diff-in-Disc methodology: properly formalizing the fuzzy design into the Diff-in-Disc framework could account for situations where treatment assignment is not strictly binary or where there is partial compliance with the treatment conditions. This extension would allow for more realistic modeling of scenarios and further case applications where policy implementation is uneven or subject to exceptions. Another valuable extension would incorporate a time-variant running variable (X_{it}) into the model and would be particularly useful in contexts where the running variable is not static over time, allowing the analysis to capture the dynamic nature of policy impacts as the relevant conditions evolve.

Finally, to further strengthen the Diff-in-Disc methodology, particularly concerning the derivative equivalence assumption, it is essential to incorporate validity tests that could involve verifying the assumption that the derivatives of the regression functions on either side of the cutoff are indeed equal in the neighborhood of the cutoff. This would help ensure that the Diff-in-Disc methodology is indeed suited for a particular application.

Withal, the use of econometrics to solve public policies constitutes a powerful instrument in the hands of those who make political decisions. The job of a researcher consists of trial and error. The field of inference transcends the boundaries of all sciences and provides a mathematical basis with finite, noisy, and incomplete information. It is at the intersection of numerous fields of knowledge such as information theory, statistical and econometric methods of inference, applied mathematics, complexity theory, decision analysis, basic modeling, and the philosophy of science where we can find better and more efficient ways of solving the demands made of us. The scientific method of working on assumptions, developing a logical chain of results, and then testing them in real-world settings is common to all of these fields and constitutes the job of a researcher as in physics, biology, or economics. Without forgetting that, as an economist, when it comes to formulating human behavior, the irrationality that Fama spoke of bites us on the heels.

References

- Abadie, A., & Cattaneo, M. D. (2018). “Econometric methods for program evaluation”. *Annual Review of Economics*. <https://doi.org/10.1146/ANNUREV-ECONOMICS-080217-053402>
- Autor, D. H., Palmer, C. J., & Pathak, P. A. (2014). “Housing market spillovers: Evidence from the end of rent control in cambridge, massachusetts”. *Journal of Political Economy*, 122(3), 661–717.
- Baskaran, T., & Lopes da Fonseca, M. (2016). “Electoral thresholds and political representation”. *Public Choice*, 169(1), 117–136. <https://doi.org/10.1007/s11127-016-0378-8>
- Bhalotra, S., Clots-Figueras, I., & Iyer, L. (2018). “Pathbreakers? women’s electoral success and future political participation”. *Economic Journal*, 128(613), 1844–1878. <https://doi.org/10.1111/ecoj.12492>
- Brockmann, M., Gasser, T., & Herrmann, E. (1993). “Locally adaptive bandwidth choice for kernel regression estimators”. *Journal of the American Statistical Association*, 88, 1302–1309. <https://doi.org/10.1080/01621459.1993.10476411>
- Butts, K. (2023). “Geographic difference-in-discontinuities”. *Applied Economics Letters*, 30(5), 615–619. <https://doi.org/10.1080/13504851.2021.2005236>
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). “Robust nonparametric confidence intervals for regression-discontinuity designs”. *Econometrica*, 82, 2295–2326. <https://doi.org/10.3982/ECTA11757>
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). “Rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs”. *The R Journal*, 7(1), 38–51. <https://journal.r-project.org/archive/2015/RJ-2015-004/RJ-2015-004.pdf>
- Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

- Cellini, S. R., Ferreira, F., & Rothstein, J. (2010). “The value of school facility investments: Evidence from a dynamic regression discontinuity design”. *The Quarterly Journal of Economics*, 125(1), 215–261.
- Chicoine, L. E. (2017). “Homicides in mexico and the expiration of the u.s. federal assault weapons ban: A difference-in-discontinuities approach”. *Journal of Economic Geography*, 17(4), 825–856. <https://doi.org/10.1093/jeg/lbw031>
- Défossez, A., Bottou, L., Bach, F., & Usunier, N. (2020). “A simple convergence proof of adam and adagrad”. *arXiv preprint arXiv:2003.02395*.
- DesJardins, S., & McCall, B. (2008). *The impact of the gates millennium scholars program on the retention, college finance- and work-related choices, and future educational aspirations of low-income minority students* [Unpublished Manuscript].
- Duchi, J., Hazan, E., & Singer, Y. (2011). “Adaptive subgradient methods for online learning and stochastic optimization”. *Journal of Machine Learning Research*, 12(Jul), 2121–2159.
- Fan, J., & Gijbels, I. (1992). “Variable bandwidth and local linear regression smoothers”. *The Annals of Statistics*, 20(4), 2008–2036.
- Frölich, M. (2007). “Nonparametric iv estimation of local average treatment effects with covariates”. *Journal of Econometrics*, 139(1), 35–75.
- Grembi, V., Nannicini, T., & Troiano, U. (2016). “Do fiscal rules matter?” *American Economic Journal: Applied Economics*, 8(3), 1–30. <https://doi.org/10.1257/app.20150076>
- Gyourko, J., & Linneman, P. (1989). “Equity and efficiency aspects of rent control: An empirical study of new york city”. *Journal of Urban Economics*, 26(1), 54–74.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). “Identification and estimation of treatment effects with a regression-discontinuity design”. *Econometrica*, 69(1), 201–209.
- Härdle, W. (1992). *Applied nonparametric regression*. Cambridge University Press.

- Imbens, G. W., & Kalyanaraman, K. (2012). “Optimal bandwidth choice for the regression discontinuity estimator”. *The Review of Economic Studies*, 79(3), 933–959.
- Imbens, G. W., & Lemieux, T. (2008). “Regression discontinuity designs: A guide to practice”. *Journal of Econometrics*, 142(2), 615–635.
- Keele, L. J., & Titiunik, R. (2015). “Geographic boundaries as regression discontinuities”. *Political Analysis*, 23(1), 127–155. <https://doi.org/10.1093/pan/mpu014>
- Lalive, R. (2008). “How do extended benefits affect unemployment duration? a regression discontinuity approach”. *Journal of Econometrics*, 142(2), 785–806.
- Lee, D. S., & Lemieux, T. (2010). “Regression discontinuity designs in economics”. *Journal of Economic Literature*, 48(2), 281–355.
- Lemieux, T., & Milligan, K. (2008). “Incentive effects of social assistance: A regression discontinuity approach”. *Journal of Econometrics*, 142(2), 807–828.
- Leonardi, M., & Pica, G. (2013). “Who pays for it? the heterogeneous wage effects of employment protection legislation”. *The Economic Journal*, 123(573), 1236–1278.
- Li, K.-C. (1987). “Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: Discrete index set”. *Annals of Statistics*, 15, 958–975.
- Lippmann, Q. (2018). “Les politiques de quotas en faveur des femmes ont-elles brisé ou surélevé le plafond de verre ?” *Revue économique*, 69(5), 849–867. https://econpapers.repec.org/RePEc:caj:recosp:reco_695_0849
- Ludwig, J., & Miller, D. L. (2007). “Does head start improve children’s life chances? evidence from a regression discontinuity design”. *Quarterly Journal of Economics*, 122(1), 159–208.
- McCrory, J. (2008). “Manipulation of the running variable in the regression discontinuity design: A density test”. *Journal of Econometrics*, 142(2), 698–714.
- Porter, J. (2003). *Estimation in the regression discontinuity model* [Unpublished manuscript, Department of Economics, University of Wisconsin, Madison].

- Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". *Journal of Educational Psychology*, 66(5), 688–701. <https://doi.org/10.1037/h0037350>
- Schucany, W. R. (1995). "Adaptive bandwidth choice for kernel regression". *Journal of the American Statistical Association*, 90, 535–540. <https://doi.org/10.1080/01621459.1995.10476545>
- Thistlewaite, D. L., & Campbell, D. T. (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment". *Journal of Educational Psychology*, 51, 309–317.
- Turner, B., & Malpezzi, S. (2003). "A review of empirical evidence on the costs and benefits of rent control". *Swedish Economic Policy Review*, 10, 11–56.
- Van der Klaauw, W. (2008). "Regression-discontinuity analysis: A survey of recent developments in economics". *Labour*, 22(2), 219–245.
- Wand, M., & Jones, M. (1994). *Kernel smoothing*. Chapman; Hall.

Appendix

Statistic	Inside Paris	Outside Paris
N	78,270.00	38,483.00
Mean	1,848.02	910.26
Median	1,679.17	726.30
SD	985.27	794.05
Min	17.01	0.03
Max	4,634.66	5,571.90

Table 8: Descriptive statistics for distance to the border from properties inside and outside Paris

Table 10: Minimal and Maximal Distances to the Border for all Paris Districts and Surrounding Communes

Commune	Min Border Distance(m)	Max Border Distance(m)
Montrouge	0.03	1396.30
Neuilly-sur-Seine	0.41	1719.00
Boulogne-Billancourt	0.54	1962.57
Vincennes	1.05	1078.78
Nogent-sur-Marne	1.11	2130.72
Saint-Mandé	1.16	395.46
Issy-les-Moulineaux	1.42	2109.33
Gentilly	1.60	776.38
Charenton-le-Pont	2.24	829.92
Le Pré-Saint-Gervais	3.83	718.02
Levallois-Perret	4.40	1425.84
Saint-Maurice	5.88	663.87
Paris 19e Arrondissement	17.01	2231.43
Paris 17e Arrondissement	18.08	1855.90
Les Lilas	23.41	1376.09
Bagnolet	27.08	1743.86
Aubervilliers	32.49	2460.26
Clichy	37.50	1452.08
Ivry-sur-Seine	38.38	2429.91
Pantin	39.58	2302.35
Le Kremlin-Bicêtre	39.71	1462.40
Montreuil	39.77	3795.72
Paris 16e Arrondissement	52.15	2617.89
Puteaux	54.18	1694.80
Vanves	54.61	1638.37

Continued on next page

Table 10 – continued from previous page

Commune	Min Border Distance(m)	Max Border Distance(m)
Fontenay-sous-Bois	62.73	2941.21
Joinville-le-Pont	66.01	1480.13
Malakoff	69.17	1989.19
Paris 18e Arrondissement	107.38	2106.60
Saint-Cloud	156.42	2192.68
Suresnes	195.63	1880.51
Saint-Denis	207.59	5571.90
Paris 15e Arrondissement	212.14	2799.46
Paris 20e Arrondissement	253.83	2202.07
Paris 13e Arrondissement	269.41	2362.12
Paris 14e Arrondissement	316.06	2445.04
Paris 12e Arrondissement	330.01	3234.89
Paris 8e Arrondissement	1262.14	3224.12
Paris 11e Arrondissement	1280.57	3442.16
Paris 9e Arrondissement	1900.83	3370.93
Paris 10e Arrondissement	1939.47	3585.29
Paris 5e Arrondissement	2110.46	3936.72
Paris 7e Arrondissement	2325.49	4243.56
Paris 6e Arrondissement	2469.24	4223.29
Paris 4e Arrondissement	3020.15	4603.54
Paris 1er Arrondissement	3236.31	4634.66
Paris 3e Arrondissement	3285.26	4359.36
Paris 2e Arrondissement	3306.91	4211.60

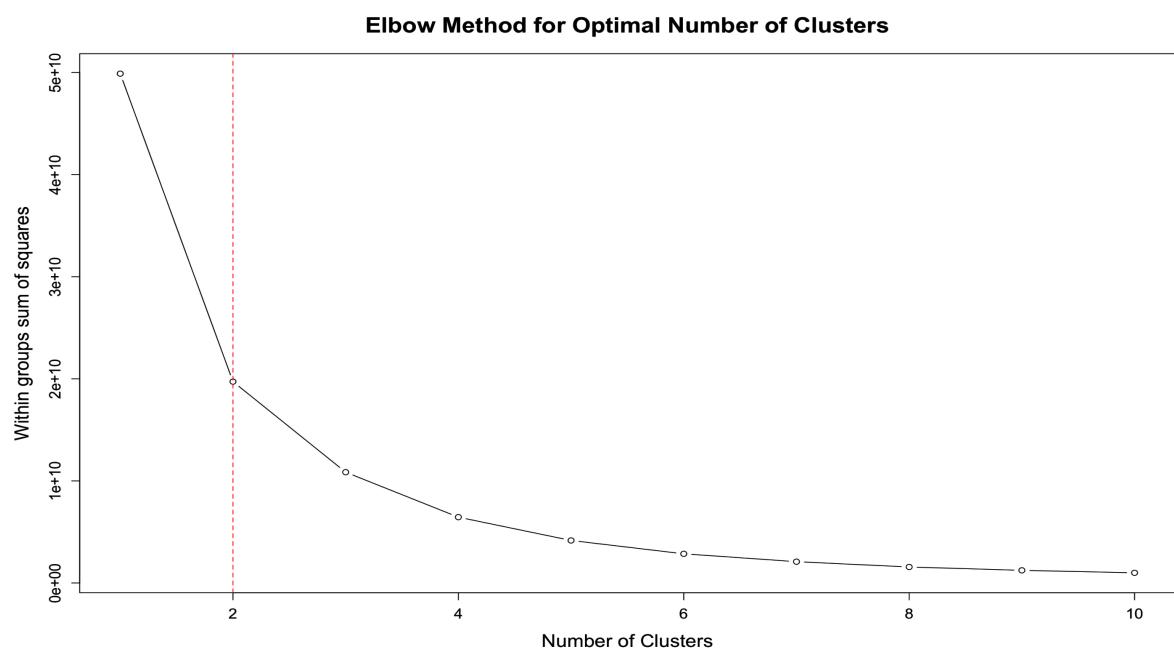


Figure 8: Elbow method for obtaining the number of Clusters K

Commune	N	Median ($\text{€}/\text{sqm}$)	SD ($\text{€}/\text{sqm}$)	Q1 ($\text{€}/\text{sqm}$)	Q3 ($\text{€}/\text{sqm}$)
Neuilly-sur-Seine	2355	10932.20	3100.37	9423.08	12381.39
Levallois-Perret	2586	9559.41	2393.36	8255.81	10543.57
Vincennes	2100	8850.29	2555.41	7500.00	10000.00
Saint-Mandé	859	8833.33	2474.27	7715.17	9912.00
Boulogne-Billanc.	4745	8813.56	2385.92	7702.70	9915.73
Issy-les-Moul.	1910	7993.99	2102.23	6734.78	8947.37
Charenton-le-Pont	944	7816.51	1985.89	6646.16	8775.83
Montrouge	1585	7457.58	2063.04	6275.86	8523.81
Puteaux	1511	7227.54	1963.77	6132.77	8139.21
Suresnes	1356	6962.50	1929.61	5988.54	7773.39
Clichy	1719	6945.06	2258.63	5658.81	7960.36
Malakoff	580	6756.76	2156.19	5793.14	7590.82
Les Lilas	597	6750.00	1871.71	5750.00	7685.19
Vanves	828	6718.04	1740.05	5927.20	7594.44
Saint-Cloud	912	6713.17	1848.74	5932.33	7641.26
Le Pré-St.-Gervais	440	6295.65	1871.66	5284.96	7072.72
Nogent-sur-Marne	1383	6171.88	1812.33	5300.00	7007.46
Montreuil	2460	6166.67	2273.28	4601.11	7371.71
Saint-Maurice	485	6052.63	1946.62	5097.37	7162.16
Pantin	1399	5896.55	1936.98	4642.86	6914.24
Le Kremlin-Bicêtre	529	5813.95	1537.14	4992.65	6500.00
Gentilly	280	5732.12	1726.27	5087.98	6668.61
Fontenay-sous-Bois	1147	5703.13	1890.83	4429.48	6756.16
Joinville-le-Pont	527	5441.18	1490.86	4721.12	6204.17
Bagnolet	675	5205.13	1692.68	4073.43	6244.59
Ivry-sur-Seine	1150	4949.19	1548.69	4167.34	5760.05
Saint-Denis	2006	3893.26	1561.64	3123.86	4576.14
Aubervilliers	1415	3666.67	1747.54	2934.19	4446.43

Table 9: Descriptive Stats for the Surrounding Municipalities