NOVA SCHOOL OF BUSINESS & ECONOMICS

University Nova de Lisboa

Dissertation, presented as part of the requirements for the Degree of Doctor of Philosophy in Economics

Data-Driven Insights in Real Estate: Data, Policy, and Shocks in Portugal's Urban Landscape

Bernardo Forbes Costa

Student nº 14652

A Dissertation carried out on the Ph.D. in Economics, under the supervision of Professor Carlos Daniel Santos and Professor Qiwei Han

19th April 2024

Data-Driven Insights in Real Estate: Data, Policy and Shocks in Portugal's Urban Landscape

This thesis explores the dynamics of Portugal's real estate sector through a data-driven lens, focusing on data enhancement, the influence of short-term rentals, and the impact of external shocks like the COVID-19 pandemic. It proposes ways of enhancing real estate data policy development, quantifying short-term rentals' effect on housing stock, and employing machine learning techniques to evaluate the pandemic's toll on Airbnb. Together, these studies highlight the role of data and advanced analytics in unraveling the complexities of contemporary real estate environments, thereby enabling the formulation of more effective policies and providing insights for stakeholders.

Keywords: Real Estate Data, Short-Term Rentals, Economic Shocks, Machine Learning, Policymaking

To my father and my grandmother.

Contents

Introduction

The real estate sector is vital to Portugal's economy, impacting urban development, socio-economic growth, and citizen well-being. This thesis examines three key areas: improving data for policymaking, analyzing short-term rentals (STRs) influence on housing supply, and assessing the COVID-19 pandemic's impact on STRs.

Chapter 1 shifts to the STRs' influence on Lisbon's housing stock. By leveraging a policy shock, this study quantifies the changes in the housing supply resulting from increased STRs. The findings suggest that STRs incentivize housing stock revitalization, potentially escalating housing costs through amenity effects and property value increases.

Chapter 2 explores avenues for public institutions to enhance real estate data generation and distribution to bolster evidence-based policy. It underscores the importance of leveraging big data, Artificial Intelligence (AI), and blockchain technologies to enhance data utility, transparency, and processing efficiency. That should allow for crafting more informed policies that resonate with current economic and societal needs, particularly in addressing the housing affordability challenges.

Chapter 3 extends the counterfactual machine learning imputation methodology to incorporate endogenous covariates to evaluate the COVID-19 pandemic's impact on Lisbon's STR (i.e., peer-to-peer accommodation) market. This paper explains the pandemic's demand shock on the sector, translating into substantial revenue losses for hosts.

Chapter 1

Short-term Rentals Are in Town: How Does Housing Supply Change?

Bernardo Forbes Costa, Sofia Franco, Carlos Daniel Santos

1.1 Introduction

Home-sharing and short-term rentals (STRs) have been a trending topic in recent years, gaining significant public attention in housing policy and in the public debate arenas since web-based platforms such as Airbnb grew as the world´s go-to platform when households informally offer accommodation to visitors. Even though STRs can provide residents with an additional source of income while encouraging the rehabilitation of the housing stock and its associated economic benefits for the city as a whole (Kaplan and Nadler, 2015), critics argue that the surge in popularity of these STR platforms has also contributed to decrease housing affordability to local residents and to intensify displacement and neighborhood gentrification in most cities. As landlords switch from long-term (which serves local residents) to short-term (which serves tourists) rentals, house prices and rents in the long-term market go up because the housing supply is inelastic in the short run.[1] A growing number of academic articles has already investigated the effect of STRs on housing rents and prices (Horn and Merante, 2017; Garcia-López et al., 2020; Barron et al., 2021; Valentin, 2021; Koster et al., 2021; Franco and Santos, 2021), showing a positive causal effect of Airbnb on house prices and rental rates in the cities where it operates, raising housing costs for local residents. Moved by these housing affordability concerns, policymakers in several cities around the world have implemented stringent regulations or even banned the provision of these rental services.[2] Yet, the potential neighborhood residential investment driven by STRs remains very overlooked.

[1] Critics also argue that STRs provide unfair competition to hotels (Lee, 2016; Zervas et al., 2017; Sheppard and Udell, 2016), contribute to racialized gentrification, and, that tourists in residential buildings create negative externalities such as noise, crowded streets, congestion, and competition for scarce parking (Edelman and Luca, 2014; Lieber, 2015; Cócola-Gant, 2016; Edelman et al., 2017; Filippas and Horton, 2018).

[2] For instance, in March 2023 Portugal's government introduced a ban on new licenses for short-term rentals as part of the country's current efforts to manage the cost-of-living and housing crisis. However, in 2018 the Lisbon city council had already created the municipal regulation of *Alojamento Local* which allowed the city to limit the allocation of new short-term renting licenses through the creation of containment areas in specific civil parishes.

Whether and when redevelopment occurs depends on the total costs of redevelopment and the difference between expected rents from the current use and from the potential new use (Clapp et al., 2012; Munneke and Womack, 2015; Wheaton, 1978, 1982; Brueckner, 1980, 2000). By increasing the option value of spare capacity in residential housing (some hosts rent extra rooms; while others convert long-term rentals into short-term rentals), home-sharing raises the potential economic returns and demand for residential housing. Therefore, we would expect also that home-sharing leads to greater redevelopment in the long run, in cities where platforms like Airbnb operate, if the costs of land assembly, demolition, and rebuilding are not extremely high. To our knowledge, only two academic studies have explored whether Airbnb has a causal effect on the quantity and type of real estate investment (Bekkerman et al. (2022) for the entire US Market and Los Angeles specifically; Xu and Xu (2021) for Chicago). Both studies show that not all STRs come from the reallocation of the existing housing stock as shown in Koster et al. (2021). However, while Bekkerman et al. (2022) provide evidence that some of the STRs in Los Angeles County come from new constructions and additions (e.g., a detached structure in existing residential lots) rather than renovations, Xu and Xu (2021) find increases in the number of renovation projects but not significant changes in residential demolitions or new construction projects associated with the Airbnb market size in Chicago. Xu and Xu (2021) further show that besides residential redevelopment, home-sharing also spurs retail revitalization, with the investment effects driven mostly by commercial hosts. All in all, the scarce evidence of the effect of STRs on residential investment suggests that Airbnb and platforms alike have the potential to spur neighborhood redevelopment and urban renewal. Furthermore, the additional residential investment brought by STRs platforms can further benefit cities by providing additional tax revenues. And, because STRs are more spatially dispersed in a city than traditional lodging such as hotels, the induced residential investment may spread to a wider area beyond conventional touristy areas.

In this chapter, we study the causal effects of home-sharing on the quantity and type of real estate investment with a double-difference estimation strategy using evidence from Lisbon, Portugal, and a comprehensive dataset of building permit requests in the municipality, from 2011 to the end of 2016. We supplement these data with Airbnb listings over the same period.

Lisbon offers an interesting case to study as it is a high-value touristy European city with a relatively old building stock and scarce vacant developable land that has experienced a rapid growth in Airbnb listings following the 2014 STRs regulatory reform, Decree-Law no. 128/2014 (henceforth "Airbnb reform").[3] According to data from AirDNA in June 2015, there were already 7,011 Airbnb listings in the Lisbon market. This number in Lisbon rose to 10,547 by the end of May 2016 indicating a 15% increase overall, supported by higher tourism demand in the historic center of the city.[4] In addition, the city is characterized by hilly narrow streets, especially in its historic districts, and a number of idiosyncratic binding constraints on building supply, including height limits due to the main international gateway from Portugal (Lisbon Humberto Delgado Airport) being located 7 Km north of the city; historic preservation rules; and natural barriers on the land area due to its location along the Tagus River. In such a context, housing investment may take different forms requiring creative reuse of developable land and existing structures rather than just complete replacement. Moreover, the large heterogeneity in Airbnb listings concentration across the civil parishes of Lisbon suggests that STRs have not affected all areas of the city equally. In this study, we then take advantage of the 2014 Airbnb reform to examine how Airbnb listings concentration (which we

[3]The 2014 regulatory reform, *Decree-Law no. 128/2014*, greatly simplified the process of entering into the short-term sublet business by just requiring a free pre-registration of the property with the Portuguese Registry Office of Short-Term Rentals (RNAL) before the host (whether a large establishment or rooms in a private home that are occasionally rented) lists the property on a hosting platform or accepts guests. In 2017, more rigorous legislation was approved in Portugal to regulate short-term rental properties requiring properties posted on Airbnb, HomeAway and other platforms to be enrolled on the national tourism register. This is outside our period of analysis.

[4]According to Lisbon's tourist board, the number of visitors to Lisbon increased by 18% from 2013–2015, and the 2016 report of Airbnb for the Lisbon market stated that 30% of the guests would not have visited Lisbon or stayed longer had they not had access to Airbnb.

will denote as “Airbnb share”) has impacted new construction of residential units, additions to existing residential structures, and remodels without expansion, across Lisbon. We define Airbnb share as the total dwellings listed on Airbnb in a civil parish expressed as a decimal fraction of the total dwellings in that spatial unit of analysis (see Section 1.3 for more details).

We start our study by describing the amount of residential investment activity that took place between 2011-2016, distinguishing between new construction, remodels, and additions. We also describe the spatial patterns of these three types of real estate investment across civil parishes. Then we test for the causal effect of Airbnb share on residential investment using a difference-in-differences (DiD) framework that focuses on the difference between touristy areas (treatment group) and other civil parishes (control) (first difference), and the change that took place following the Airbnb reform in 2014Q3 (second difference). We define “touristy” areas the civil parishes with the largest Airbnb share and the pre-and-post-Airbnb expansion periods as the periods of high and low Airbnb growth. Furthermore, we run a second specification, an instrumental-variable (IV) approach, wherein in the first stage we instrument Airbnb share on the interaction between the dummies for treatment/ control and pre-/ post-reform, and in the second stage, we use the estimated Airbnb share to explain the three types of real estate investment described. This should enable us to validate the results from our primary specification while simultaneously estimating the elasticities of real estate investment in response to Airbnb concentration. Lastly, we turn our attention to the specific permit types, namely licensing, advance notice, and advance information requests, and replicate the same analyses described above. Of these, only licensing and advance notice pertain to the three real estate investments previously mentioned, namely new constructions, remodels, and additions. Meanwhile, advance information requests serve as an optional avenue for investors to obtain preliminary insights about the feasibility of certain urban projects. Our aim is to gauge investor responses to the Airbnb reform. Given that each permit type has its unique time frame implications, they provide varying degrees of immediacy and ease of approval. For a comprehensive breakdown of the real estate investments and permit types, please see Section 1.2.

Our results are consistent with the theoretical predictions from urban growth models with durable housing, indicating that the type and quantity of residential investment spurred by the increase in Airbnb share vary substantially across civil parishes within the city of Lisbon. We find that, in comparison to other parishes, touristy civil parishes had, on average, an increase of 18.4% and 38.2% for additions, and remodels, respectively, as a result of the Airbnb reform (i.e., after 2014Q2). On the other hand, the effect on new construction is negligible and not statistically significant. Moreover, investors appeared to favor quick-permitting projects, aiming to swiftly leverage the policy change and transform properties into Airbnb revenue sources. Specifically, in touristy parishes, the Airbnb reform resulted in a 41.8% average uptick in advance information requests compared to other parishes, and a 34.1% average disparity for advance notices. Conversely, licensing procedures remained unaffected (not statistically significant effect).

Our study contributes to the broader literature on the effect of short-term rentals on the housing market (Sheppard and Udell, 2016; Horn and Merante, 2017; Almagro and Domínguez-Iino, 2022; Garcia-López et al., 2020; Barron et al., 2021; Valentin, 2021; Koster et al., 2021; Franco and Santos, 2021) and in particular to the scarce but rising stream of research studying home-sharing and its effects on real estate investments (Bekkerman et al., 2022; Xu and Xu, 2021). To our knowledge, this is the first study to estimate the causal effects (using a quasi-experimental variation provided by a change in STRs regulation) of Airbnb on housing supply, namely on the supply of new housing stock and changes to the existing one, in the context of a large European city. This is important given the underlying differences between European and US cities. For instance, European cities might have less excess capacity, where guest houses or basement apartments are pretty much nonexistent, and fewer vacant developable lots in central areas. European cities also have many historic environments and buildings subject to strict planning controls, listed building consent, historic preservation rules, and conservation area controls. Furthermore, it is also not unusual to find in European cities heritage buildings (e.g., convents, palaces, churches, old shipping warehouses, and abandoned industrial buildings to name a few) for which the original use is no longer relevant or viable, offering an attractive investment opportunity for adaptive reuses.[5] While areas with low housing supply elasticity are bound to create large price effects (Saiz, 2010), housing investment in areas with relatively old building stock, scarce vacant lots, and strict restrictions on housing supply may also take different forms than new construction or teardowns. In our study, we use a DiD estimation strategy and data on building permit requests to explore how Airbnb share affects the quantity and form of housing investment across civil parishes within the high-value city of Lisbon.

Our study also relates to the literature studying the effects of land use regulation and zoning (since the Airbnb reform can be interpreted as an example of a land use change) on housing supply (Thorson, 1997; Levine, 1999; Mayer and Somerville, 2000; Pendall, 2000; Glaeser and Ward, 2009; Chakraborty et al., 2009; Jackson, 2016). Studies in this strand of literature show that strict local land use controls reduce housing construction or shift new housing developments to adjacent municipalities with less restrictive land use controls (Levine, 1999). Most of these studies compare land use regulations and housing constructions across several municipalities or metropolitan areas in the United States. An accompanied large set of studies measuring the effects of land use regulation on house prices also finds that in the United States, more restrictive regulations on new housing supply are associated with higher home prices (Mayer and Somerville, 2000; Glaeser and Gyourko, 2003;

[5]For instance, a very well-known successful example of adaptive reuse development in the city of Lisbon, Portugal, is the redevelopment of the historic waterfront area known as Cais do Sodré into a lively hub of bars, restaurants, and shops, being now one of the most popular nightlife destinations in the city. Another good example is the redevelopment of the Palace of the Counts of Ribeira Grande located in the civil parish of Alcantara which used to be a high school in the 80s and now is being reused and redeveloped to become a 5-star hotel and a museum.

Quigley and Rosenthal, 2005; Glaeser et al., 2006; Ihlanfeldt, 2007; Glaeser et al., 2008; Glaeser and Ward, 2009; Gyourko et al., 2013; Saiz, 2010; Kok et al., 2014; Hilber and Vermeulen, 2016). It should however be noted that even if a regulatory change yields a supply increase, prices may not fall (or stop rising). Recently, Büchler et al. (2021) provided evidence for the Canton of Zurich in Switzerland that local upzoning which refers to relaxing local land-use regulations on a parcel to allow for more housing construction is an effective policy for increasing the housing supply without increasing rents at the local level. The reason is that in rezoned locations, builders might convert existing lower-cost units into higher-cost ones; the amenity effects resulting from these conversions plus associated neighborhood retail and public safety improvements may, in turn, increase surrounding housing values. In contrast to past studies (except Büchler et al. (2021)) that rely on cross-sectional variation in land use regulation or an index of the local regulatory environment, our research design explores the regulatory effect on local housing supply for an individual European municipality based on a change in regulation (that has loosened land/housing use restrictions) at a specific point in time. Moreover, our detailed housing panel data allow us to study the local variation of the effect of the STRs regulatory reform not only on new construction and other types of local housing investment within the city of Lisbon but also on how these effects varied over time. Future work should explore how the additional real estate investment driven by short-term rentals impacted property values.

The rest of the chapter is organized as follows: Section 1.2 provides trends and the spatial distribution of short-term rentals and the different types of real estate investment in the municipality of Lisbon during our study period, as well as a brief description of the types of legal permissions often required for building projects and alterations, namely licensing or advance notices (mandatory) and prior information requests (optional). Section 1.3 describes our data, while Section 1.4 presents our identification strategy, and exhibits and discusses our main results. Finally, Section 1.5 offers conclusions.

1.2 Overview of the Housing Market in Lisbon

1.2.1 Building Permits in Portugal

In Portugal, construction and land development are governed by *Regime Jurídico da Urbanização and Edificação* (RJUE).[6] Municipalities are responsible for the control of urban planning projects, including the construction and use of buildings, and allotment of land. Municipalities also approve city urban planning and building regulations as well as policies regarding the payment of taxes/charges for real estate activities.

The execution of urban planning projects must generally be examined in advance and may require either (i) **licensing** (*licenciamento*) or (ii) **advance notice** (*comunicação prévia*). Urban planning projects for which advance notice must be provided are usually either less relevant in urban planning terms than those which require licensing or are performed in areas covered by an allotment operation or by a detailed plan (*plano de pormenor*).[7] Advance notice procedures are simpler and faster than licensing procedures.[8] Licenses or acceptances of advance notice generally expire if not implemented within the period established by law. Depending on the complexity of the urban project, the license approval process may take on average between 2 to 4 years, while advance notices can take up to 1 year to be approved. While not

[6] See the RJUE (in Portuguese).

[7] According to RJUE, real estate investment activities with a major impact on the housing stock require licensing such as (i) allotment operations; (ii) urbanization works in areas not covered by allotment operations; (iii) works in areas not covered by allotment operation or layout plan; (iv) works in classified/proposed buildings, as well as (v) in buildings integrated in classified/proposed assemblages or sites; (vi) works in buildings located in historic conservation areas; and (vii) works that result in the removal of tiles from the building or home façade.

[8] According to the RJUE, real estate investment activities requiring advance notice include: (i) reconstruction works which will not result in an increase in the façade's height or in the number of floors; (ii) urbanization works and renewal works in plots in an area covered by allotment operations; and (iii) construction, modification or extension works in an area covered by allotment operations or layout plans; (iv) works that derive from an order given by the municipality.

mandatory, it is also common practice that a potential real estate investor files an **advance information request** (*pedido de informação prévia*) with the municipality to get formal confirmation on the feasibility of a specific urban project as well as to any legal or regulatory constraints before submitting its respective specific permit (licensing or advance notice) request.[9] This process is much faster and should take between 20 and 30 days for the municipality to provide an answer to the request. It should be noted that the timelines mentioned pertain to approval/ response to submitted permit requests. Thus, even if there is an inherent time gap between the submission of the permit request and the commencement of the real estate activity (i.e., the issuance of the permit), our focus on requests allows us to pinpoint the exact date of the investment intention. In our chapter, while we also look at advance information requests, our primary focus is on licensing and advance notices permit requests as these involve real estate investments that directly impact the housing supply.

1.2.2 Types of Real Estate Investment

We can define three types of real estate investment activity with an impact on a city´s housing stock:

(a) **Remodels** encompass interventions in existing buildings that aim to modify certain features without expanding their area, as well as actions to maintain, restore, repair, and clean the structure, preserving its original or previous conditions. Specifically, this includes changes to the physical characteristics, structural adjustments, alterations in the number of housing units or interior divisions, and modifications to the nature and color of exterior finishes, all without increasing the total construction area, footprint, or facade height. It also covers restoration efforts, repairs, and cleaning tasks to maintain the build-

[9]For instance, a real estate investor may request the municipality to provide advance information on different features of a specific potential urban project, such as eaves height, number of apartment dwellings, or an estimate of urban planning costs/charges.

ing's original state. Remodels are somewhat synonymous with *renovations*, however, the goal of a remodel is to change the structure, façade/appearance, and function of the space, while the goal of a renovation is to repair or restore the structure.[10] Some renovations are subtle and hyper-focused on a specific element, and some renovations are more substantial and overlap with remodels. An architect will come up with plans for these projects and hire a contractor to do the construction.[11]

(b) Home **additions** occur when you're adding onto an existing building. For example, another bedroom, or a living room. These projects generally require demolition but not as much as a remodel. For additions, it is necessary to have an architect, contractor, and a building permit. It may also be necessary for an electrician, plumber, and mechanic depending on what is added to the structure.

(c) A **new construction** project is a home or building being built from the ground up. For this type of project, it is necessary to go through everything such as building permits, soil reports, surveys, etc. An architect, contractors, and subcontractors will be needed as well.

We refer to the three types of investment described as permit request "subtypes" and to the overarching categories of licensing, advance notice, and advance information requests as permit request "types". This is because these three types of investment can originate from either licensing or advance notice. Note that there are other permit subtypes that originate from licensing or advance notice (e.g., demolitions). However, advance information requests are not related to any of these

[10] Another type of project similar to a remodel is retrofitting. With this type of project, something new is added to the original building or structure, however, the goal of a retrofit is to specifically improve the functionality of the building by adding new technology, building systems, or equipment.

[11] Remodels result from grouping several subtypes of permit requests, more specifically alterations (*alteração*), conservation works (*obras de conservação*), and waving of requirements (*dispensa de requisitos*). It is noteworthy that the waiving of requirements permit request subtype was repealed in 2013 due to legislative changes (see Licenciamento Zero). However, we choose to retain it in our data because it encompasses projects that fall within the scope of remodels.

categories and represent a complementary, standalone category, as previously discussed.

1.2.3 Touristy civil parishes

Our definition of the treatment group (i.e., "touristy" parishes) was primarily influenced by the concentration of Airbnb listings in each civil parish. We propose that areas with higher Airbnb share should be the most impacted by the Airbnb reform. For a formal definition of "Airbnb share" and the respective values for each civil parish please see Section 1.3 and Table 1.1. To provide additional support to our treatment selection we collected additional data from the Lisbon Municipal Hall on the number of hotels and monuments for each civil parish (see Appendix - Table A1).

Parishes like Santa Maria Maior, Santo António, and Arroios not only have a notable Airbnb presence before the policy but also host a significant number of hotels and monuments. The concentration of monuments in parishes such as São Vicente, Misericórdia, and Estrela showcases their cultural or historical importance, marking them as intrinsic attractions for tourists. On the other hand, while Avenidas Novas does have a substantial number of hotels, its role as a business hub for the Lisbon municipality means it caters to a different type of audience than the typical users of STRs. This is further underscored by the presence of only one monument in the area. São Domingos de Benfica, despite having a few monuments, is primarily residential in nature. This likely caters to those seeking an authentic, localized experience, which might explain its lower Airbnb share and number of hotels. Belém presents a similar case: Despite its rich collection of monuments, the combined low numbers of hotels and Airbnb listings hint at its role as more of a day-visit location rather than an overnight destination for tourists.

Lastly, the legitimacy of our treatment group definition is further bolstered by subsequent policy decisions, specifically *Proposal 677/CM/2018*, from November

2018, by the Municipal Hall of Lisbon. This policy, even though it was implemented outside of our period of analysis, officially recognized the very parishes we had identified as "touristy" as the most impacted by Airbnb share growth, underlining the enduring applicability of our initial criteria.

Next, we overview the housing market in the municipality of Lisbon, examining the trends and spatial distribution of short-term rentals and types of real estate investment (new construction, remodels, and home additions), between 2011 and 2016, while also looking at permit types as a whole (licensing, advance notice, advance information requests).

1.2.4 Trends and Spatial Distribution of STRs and Real Estate Investment Types

Figure 1.1 displays the number of STRs listings before and after the Airbnb reform, along with the respective means for the two periods. The data shows a significant increase in the number of new listings during the post-policy period. To further explore the spatial implications of the policy for the short-term rental market, Figures 1.2 and 1.3 present Airbnb share by civil parish over time. Figure 1.2 suggests that prior to the policy change, none of the civil parishes exceeded a 20% Airbnb share. However, after the policy reform, there was a noticeable increase in Airbnb listings, especially in touristy civil parishes. In addition, Figure 1.3 illustrates a diminishing growth in Airbnb share as the distance from touristy parishes increases. For instance, Santa Clara had a 0.02% Airbnb share before the policy reform and a 0.28% Airbnb share at the end of our study period. On the other hand, Santa Maria Maior, a very touristy civil parish, had a more pronounced change in Airbnb share after the policy reform exhibiting a pre-reform 10.35% Airbnb share and a post-reform Airbnb share of 39.91% at the end of our study period. This indicates that there was not only a surge in Airbnb share following the policy but also underscores that this rise was particularly marked in the touristy civil parishes.

Figure 1.1: New Airbnb listings

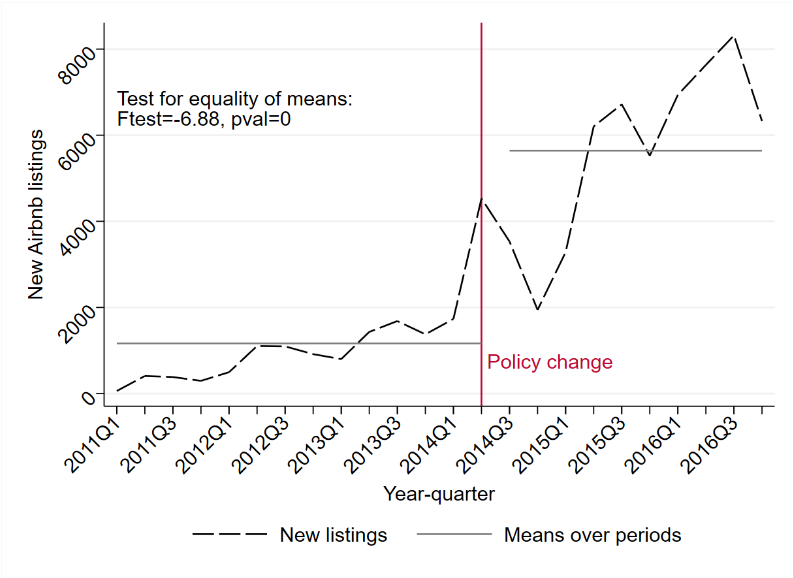
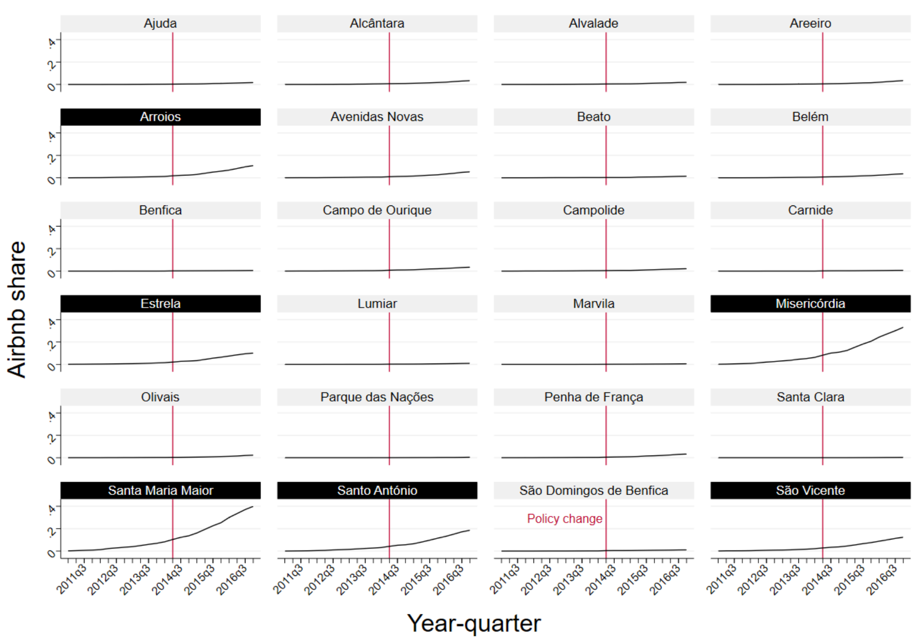


Figure 1.2: Airbnb share by parish - Time series



Notes: Graphs by civil parish. Civil parishes marked in black correspond to the touristy ones.

Figure 1.3: Airbnb share by parish - Spatial distribution

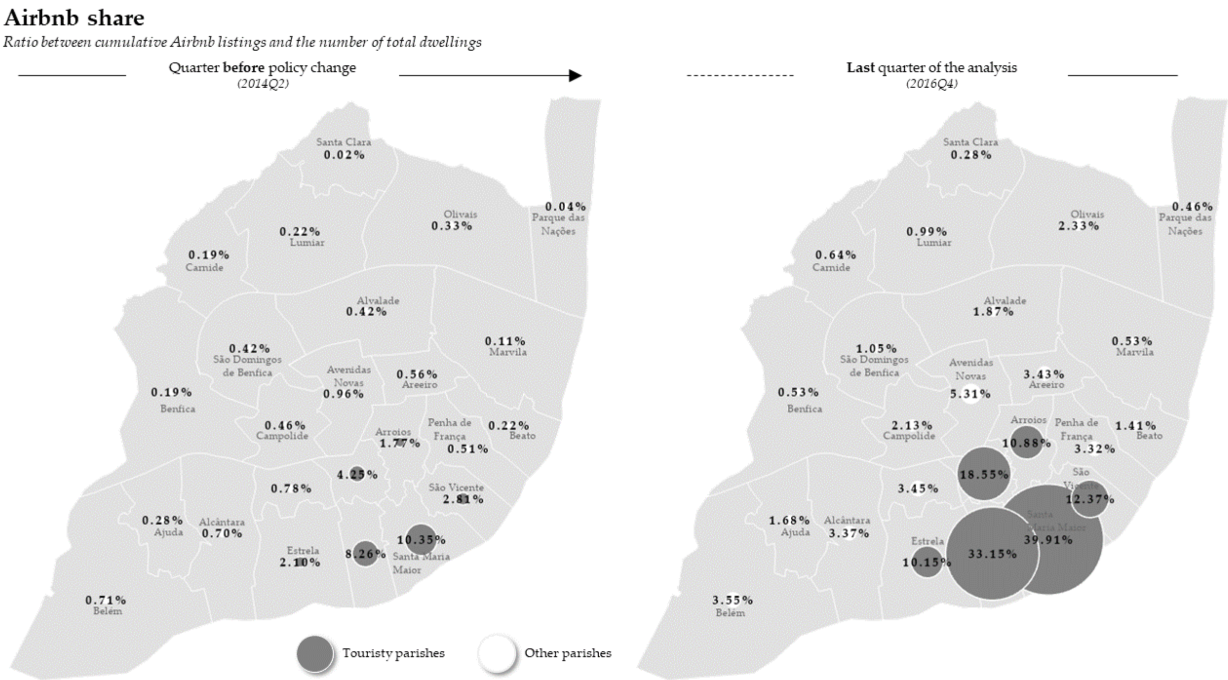


Figure 1.4: Total permit requests

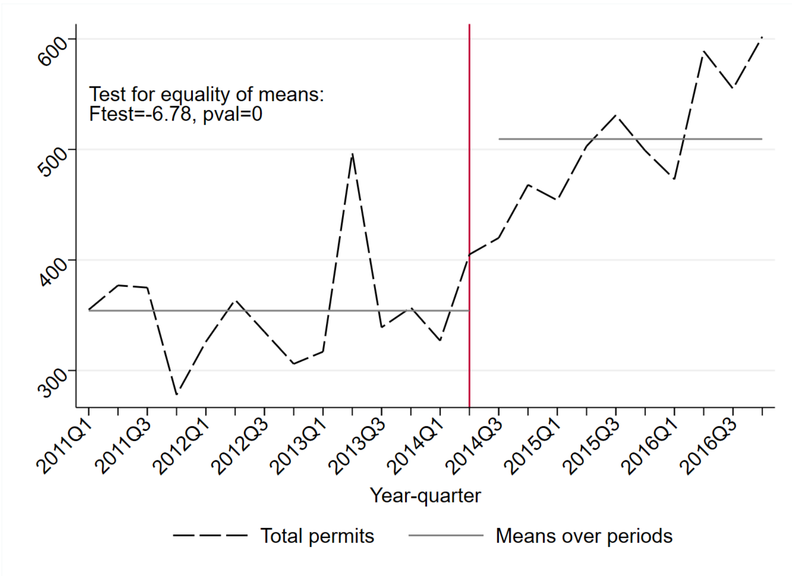
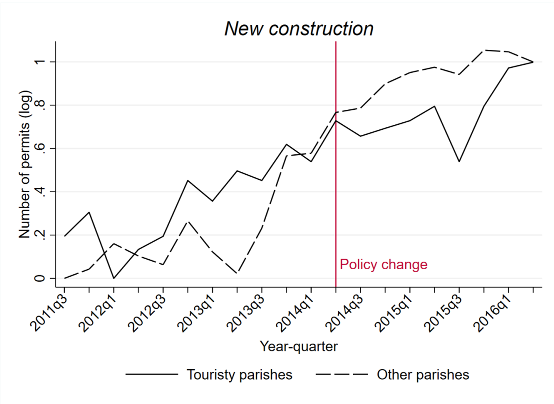


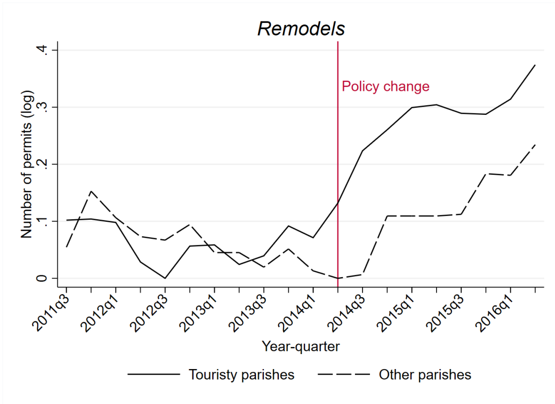
Figure 1.4 shows the total number of building permit requests (licenses and advance notices) across different types of real estate investment projects (new construction, additions, and remodels). We plot the moving average of the natural logarithm of permit requests for touristy and non-touristy civil parishes to visualize the effect of the policy reform on the housing supply. From Figure 1.4 we observe that there was a more pronounced general increase in permit requests in touristy relative to non-touristy civil parishes after the policy reform. In Figure 1.5, we can see that this happens in particular for investment projects related to additions and remodels, whereas for new construction the difference between touristy and non-touristy civil parishes is negligible (and perhaps in the opposite direction). Such an investment pattern is well aligned with the general lack of vacant land for new construction throughout the city and with the significant clustering of derelict buildings with important historic amenities in the city center. Furthermore, there appears to be an abrupt decline in advance notice requests for the non-touristy civil parishes immediately after the reform, whereas the numbers remain relatively steady for the touristy ones. Finally, we observe a rise in advance information requests specifically for the touristy civil parishes. These latter observations collectively propose that post-reform, the two types of permit requests characterized by shorter implementation timeframes—advance notice and advance information requests—underwent a significant shift towards the touristy civil parishes compared to the other ones. This suggests that investors decided to prioritize faster and less demanding permitting projects, driven by the desire to promptly capitalize on the policy implementation, and convert their properties or acquire new ones to convert them into income-generating Airbnb listings.

Figure 1.6 plots the spatial distribution of the different types of real estate investment, before and after the policy. Consistent with the patterns identified in Figure 1.5, these maps provide visual evidence of a substantial upswing in the number of additions and remodels in touristy civil parishes subsequent to the policy

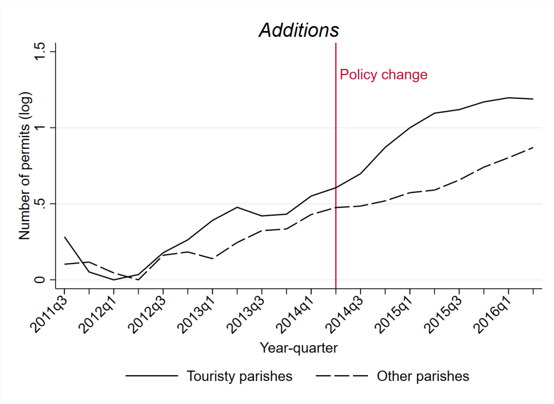
Figure 1.5: Number of permit requests



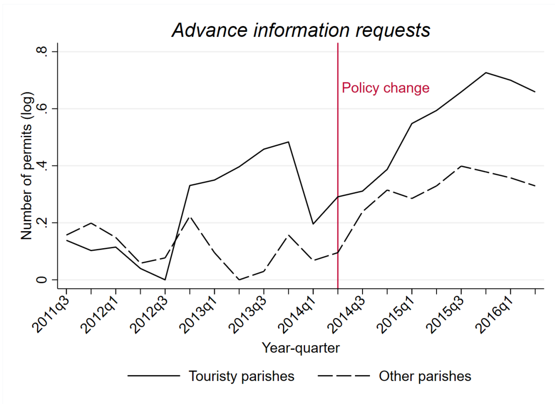
(a)



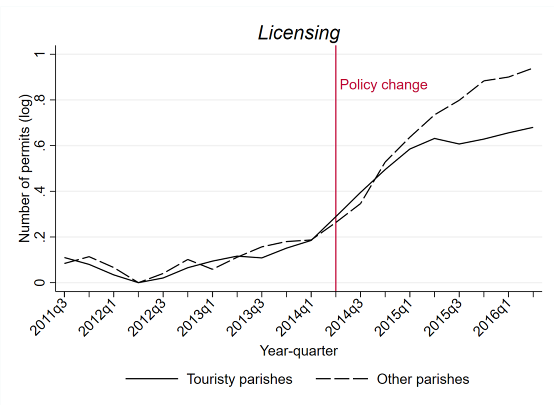
(b)



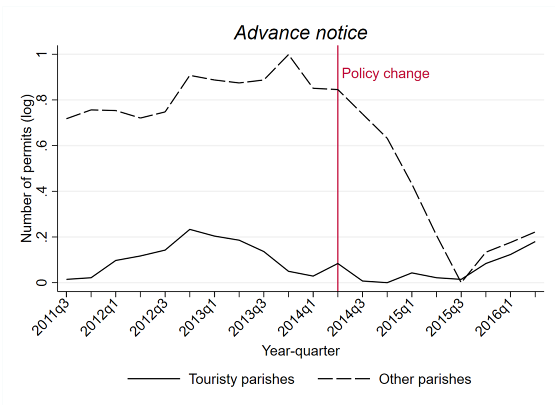
(c)



(d)



(e)

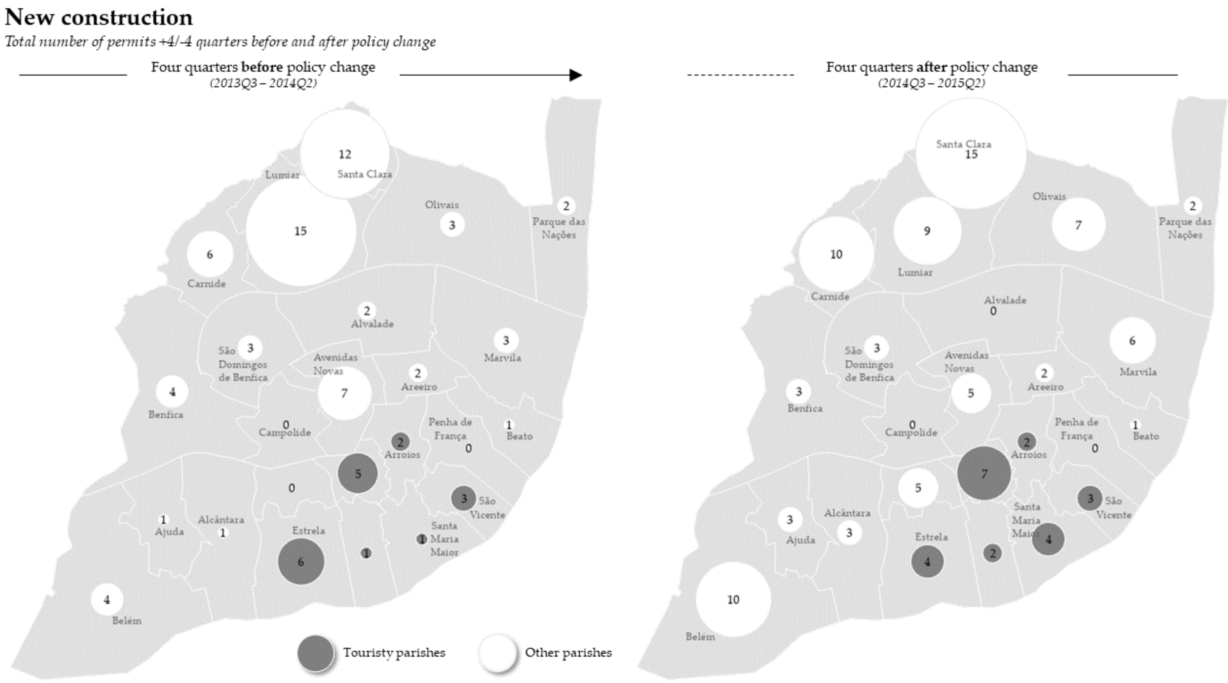


(f)

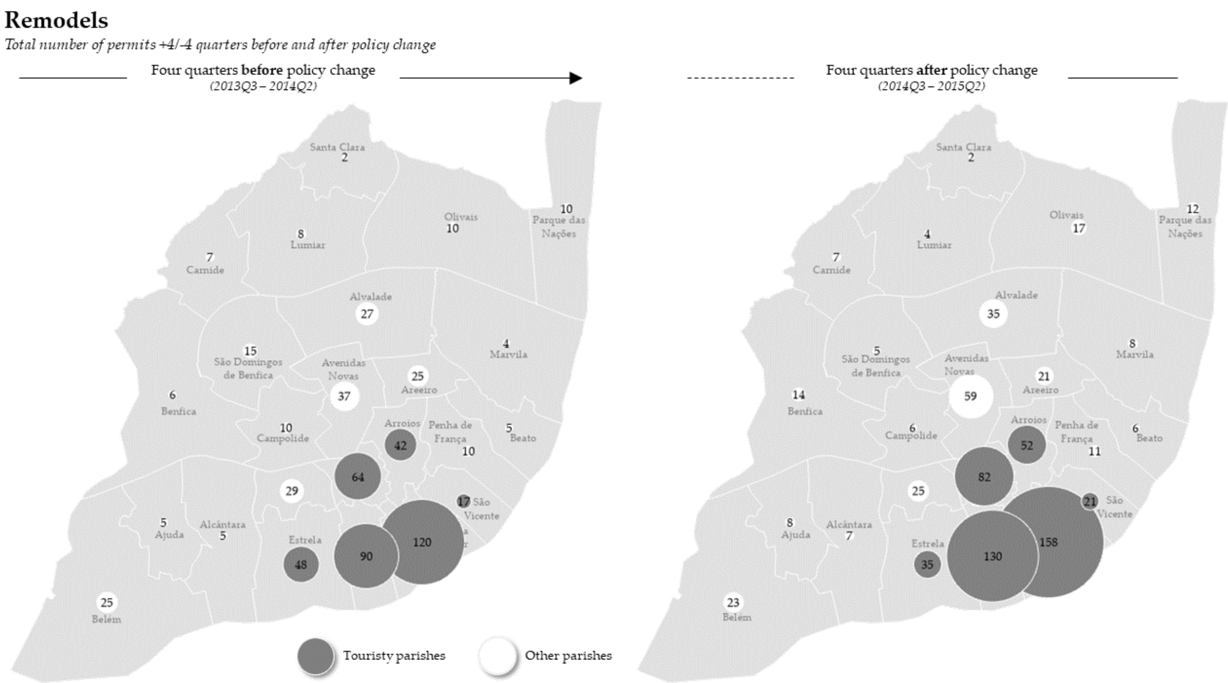
Note: Number of permits calculated as a moving average with -2/+2 quarters before and after. Each series is deducted its respective minimum value (a constant) to reduce the vertical space between them. Panels (a) - (c) refer to investment types (i.e., permit subtypes), while (d) - (f) represent permit types.

change. We also observe that remodels are much more spatially concentrated than additions, specifically in the downtown/ historic areas where most derelict buildings are concentrated. Furthermore, the spatial patterns of housing additions induced by Airbnb share can be explained by the standard urban redevelopment theory: this type of investment is more prevalent in neighborhoods with higher pre-existing housing values and older housing which also include neighborhoods adjacent and in close proximity to historic areas. Finally, the number of new construction requests does not exhibit a significant response to the reform and seems to be more pronounced on the outskirts of the city, on the opposite end of the touristy area.

Figure 1.6: Spatial distribution of investment types



(a)



(b)

Figure 1.6: Spatial distribution of investment types (continued)

(c)

1.2.5 Real Estate Investment in the Residential Market

Real estate investment can cover both residential buildings and non-residential buildings, suggesting that we can also have residential (covering single-family houses and multi-familify buildings) and non-residential permits (covering properties such as retail, offices, and industrial). However, the municipality of Lisbon does not provide information on the permit type regarding the ultimate use. As a result, we cannot separate residential from non-residential permit requests. For this reason, we call on auxiliary data on annual *issued* permits for the municipality of Lisbon data from PORDATA, a certified statistical database about Portugal, its municipalities, and Europe. The data shows that out of all the issued permits during our study period, around 80% are classified as residential (see Appendix — Table A2). This in turn suggests that the majority of the real estate investment decisions refer to the residential market. Note that these are permits that have already been issued and therefore do not directly overlap with our data.

In conclusion, the findings presented in this section offer valuable insights into the housing market in Lisbon. The data indicate that the Airbnb reform exerted a discernible impact on the short-term rental market and housing supply dynamics, particularly in parishes characterized by high tourist activity. The anecdotal evidence suggests that policy intervention facilitated a surge in Airbnb's presence and triggered rehabilitation efforts in these areas. However, the influence on new construction activities remained limited due to physical space constraints and existing legislative frameworks. Moreover, it appears that investors favored expedited permitting processes to promptly leverage the benefits of policy implementation. Building upon these observations, Section 1.4 will present the main empirical specification, leveraging the insights gleaned from this analysis of the Lisbon housing market.

1.3 Data

We collected data from Airbnb, the Municipal Hall of Lisbon, and the National Institute of Statistics for the municipality, with details for each of its 24 civil parishes.[12] [13] Civil parishes are geographic units defined at a more aggregate level than census sections and subsections.

As in many European countries, Airbnb is not the only short-term rental platform active in Portugal. However, it is the largest player in the market.[14] Furthermore, most STRs are advertised through more than one platform. This makes it highly likely that most properties will be listed on the Airbnb platform. Hence, we consider Airbnb listings a good proxy for the STR market in Portugal. In the following

[12] Airbnb is a peer-to-peer short-term rental accommodations platform, which allows owners to list their properties for rental (constrained by local regulation) and guests to book these properties for short periods of time (typically a few days or weeks).

[13] Civil parishes comprising the municipality of Lisbon: Ajuda, Alcântara, Beato, Benfica, Campolide, Carnide, Lumiar, Marvila, Olivais, São Domingos de Benfica, Alvalade, Areeiro, Arroios, Avenidas Novas, Belém, Campo de Ourique, Estrela, Misericórdia, Parque das Nações, Penha de França, Santa Clara, Santa Maria Maior, Santo António, São Vicente.

[14] As of now, according to AirDNA 95% of the STR properties in the Lisbon municipality are listed on Airbnb. For the entire country, the number stands at 73%.

subsections, we discuss these data sources and present some descriptive statistics for the main variables considered.

We build our analysis on the following datasets for the municipality of Lisbon:

1. **Real-estate permits**. Real estate permit requests, between 2009 and 2019. The data comprised the following features: permit ID, permit request type (procedure), permit request subtype (urbanistic operation), date of permit request, and location (civil parish). As previously discussed, in terms of permit types, we focus on licensing and advance notices, as these involve direct real estate investments, but also look at advance information requests. Furthermore, following Schuetz (2020), we filter the permit request subtypes to consider only (i) additions to or amplification of existing structures, (ii) remodels of existing structures without expansions, and (iii) construction of new residential units; we refer to these subtypes of permit requests as "additions", "remodels", and "new construction".[15] However, it is important to note that when examining permit types, we also consider other permit request subtypes (i.e., other urbanistic operations, such as demolitions), that are not encompassed by the three investment types previously mentioned. In other words, the sum of licensing and advance notice exceeds the sum of remodels, additions, and new construction permit requests.

2. **Airbnb listings**. Information on Airbnb listings from AirDNA.[16] It contains a total of approximately 53,000 listings and it includes the following features for each property: location (civil parish), physical characteristics (e.g., number of rooms, number of bathrooms), host id, the list of all reviews from guests who have stayed at the property, and the year-month in which the user registered as a host on the platform, with registration dates ranging from 2008

[15] Our dataset included information about redevelopment permits, composed of demolition of existing structures (*demolição autónoma* and *demolição*) and reconstruction (*reconstrução*). However, we chose not to include these as the number of observations limited our ability to conduct statistical analysis.

[16] airdna.co/about

to the beginning of 2022. In 2013 there was a reorganization in the administrative boundaries of the municipality of Lisbon and its 53 civil parishes were converted to 24. Given the location information provided by AirDNA was using the old convention of civil parishes, we converted them to the new denominations.

3. **Census 2011 and 2021**. Collected the variable "total number of dwellings per civil parish" for both censuses, which are run each decade by the National Institute of Statistics. We use the data collected from the two censuses to compute a linear interpolation between them to get a yearly estimate for our time horizon. The main purpose of this information/ variable is to compute the concentration of Airbnb ("Airbnb share") in each civil parish, as described below.

Airbnb share. First, we use the information about Airbnb listings to define an entry or a listing on the platform, by quarter. We consider an entry takes place in the quarter in which the host registers the property on the platform. However, about 17% of the listings do not have information about the registration date. Furthermore, the number of listings registered (i.e., new listings) from 2008 to 2010 is negligible given that Airbnb was not well established in the Lisbon market. Second, we define "Airbnb share" as the concentration of listed Airbnb properties in a certain geographical area (civil parish), given by the ratio of cumulative Airbnb entries to the total number of dwellings (collected from the Census) in that year. Third, we define as "touristy" and as the treatment group the civil parishes with the highest Airbnb share prior to the policy implementation, as these were most susceptible to the effects of the Airbnb reform (see Table 1.1 and Appendix – Table A3 for a comparison between treatment and control groups).[17]

Housing supply. Our main dependent variable is housing supply at the quar-

[17] As a robustness check in Section 1.4, we test changing the composition of the treatment group to consider "historical" civil parishes instead and show that the effects disappear.

terly level. We consider its quantity and quality to be given as a proxy by (a) the number and (b) the subtype of permit request (i.e., remodels, additions, or new construction), respectively. We focus on the period from 2011 to 2016 for the following reasons: (i) Airbnb was only firmly established in Lisbon by 2011, (ii) to ensure our results are not influenced by a 2018 policy intervention (and its anticipated effects in 2017) or the COVID pandemic, which emerged in 2019, and (iii) the availability of Census data is limited to 2011 onwards, allowing us to only compute the Airbnb share from that point forward. Finally, we winsorize the number of permits requests at the civil parish level, using 1% and 99% percentiles to remove outliers in our data. This reduces our dataset to 7,345 permit requests for the three types of investment considered (see Table 1.1).

As discussed, we aggregate the datasets mentioned above at the quarterly and civil parish levels, focusing our analysis on the period from January 2011 to December 2016. For a detailed description of our final data set please refer to Table 1.1. We note that we find a similar distribution of types of permits to Schuetz (2020), with remodels being the dominant category (57%), followed by addition (35%), which is also significantly larger than the new construction (7%):

> "The vast majority of permits were for alterations of existing structures (72%). Nearly 10 percent were for additions (expansions) to existing structures. New buildings represented only four percent of construction permits (. . .)." – *Schuetz, 2020*

1.4 Empirical Strategy and Results

We take advantage of a regulation change that facilitated the listing of properties as STR, the Airbnb reform (*Decree-Law no. 128/2014*), to estimate the impact of Airbnb on the housing supply.

Table 1.1: Summary statistics

Civil parish	**Touristy parish**	**Airbnb share**	**Licensing**	**Advance notice**	**Advance info. requests**	**New construction**	**Remodels**	**Additions**
Ajuda	-	0.28%	60	132	32	16	44	101
Alcântara	-	0.70%	29	138	30	13	57	68
Alvalade	-	0.42%	139	209	53	11	185	119
Areeiro	-	0.56%	111	111	24	9	135	58
Arroios	Yes	1.77%	212	429	86	25	307	213
Avenidas Novas	-	0.96%	169	399	92	28	278	168
Beato	-	0.22%	47	58	17	8	39	48
Belém	-	0.71%	131	426	51	45	157	231
Benfica	-	0.19%	69	103	23	17	63	81
Campo de Ourique	-	0.78%	138	238	51	16	168	139
Campolide	-	0.46%	36	114	35	8	46	71
Carnide	-	0.19%	32	110	17	36	40	38
Estrela	Yes	2.10%	121	519	99	34	282	210
Lumiar	-	0.22%	62	97	24	43	45	39
Marvila	-	0.11%	106	71	28	25	37	37
Misericórdia	Yes	8.26%	97	939	70	11	651	213
Olivais	-	0.33%	116	205	45	38	101	135
Parque das Nações	-	0.04%	44	50	16	15	49	20
Penha de França	-	0.51%	82	88	29	4	83	59
Santa Clara	-	0.03%	30	117	6	65	12	16
Santa Maria Maior	Yes	10.35%	108	1186	97	9	828	187
Santo António	Yes	4.25%	113	694	107	28	422	194
São Domingos de Benfica	-	0.42%	78	98	22	19	75	51
São Vicente	Yes	2.81%	76	186	50	17	112	93
TOTAL			2,206	6,717	1,104	540	4,216	2,589

Note: Touristy civil parishes correspond to the ones with the largest Airbnb share (in 2014Q2, before the policy change), given by the ratio of cumulative Airbnb listings to the total number of dwellings. The number of listings for each permit type (i.e., advance information requests, licensing, and advance notice) or permit sub-type (i.e., new construction, additions, and remodels) is the total number of permits requested in our period of analysis (2011Q1 - 2016Q4) for each civil parish.

1.4.1 Baseline OLS Specification

We let Y_{it} denote the natural log of permits in civil parish i, in the Lisbon municipality, in period t (in a given quarter-year), $AbbShare_{it}$ the share of listed Airbnb properties to the total dwellings in a civil parish i in period t. τ_t and μ_i represent year-quarter and civil parish fixed effects, respectively. Finally, we let ε_{it} contain civil parish level shocks to real estate permits. Our baseline specification is thus represented as follows:

$$Y_{it} = \alpha_0 + \beta_1 AbbShare_{it} + \tau_t + \mu_i + \varepsilon_{it} \tag{1.1}$$

where β_1 is the parameter of interest and standard errors are clustered at the civil parish level to account for serial correlation within panel units.

Table 1.2: OLS estimates of the effect of the Airbnb share on investment types

Dependent variable:	N. permit requests (log)		
Investment type:	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)
Airbnb share	0.859	1.020*	2.160***
	(0.825)	(0.534)	(0.588)
Quarter-year FE	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes
Observations	287	522	513
R-squared	0.371	0.793	0.641

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Simple OLS results with fixed effects at the civil parish and quarter-year level.

Our baseline OLS results will most likely be biased due to endogeneity issues, namely due to unobserved civil parish-specific, time-varying factors contained in the error term which are, in turn, correlated with Airbnb share. Note that we include time, as well as civil parish level fixed effects. Nevertheless, we can still see from Table 1.2 that there seems to be a correlation between Airbnb share and certain types of real estate investment. More specifically, a one percentage point (1pp) increase in a civil parish Airbnb share is associated with a 1.02% and 2.16% increase, on average, in the number of remodels and additions, respectively. However, this correlation does not seem to exist for new construction – given the lower and not statistically significant coefficient. This adds to the previous descriptive evidence, from Section 1.2, that Airbnb concentration seems to be connected to the requalification of the existing housing stock, rather than to the supply of new one. To overcome our identification issues and obtain a valid causal effect, in the next section, we turn to our main specification a difference-in-differences approach. We then complement these results with an instrumental-variable regression to help us validate the results from our main specification and to clarify the effects come from the reform and the implicit Airbnb expansion.

1.4.2 Estimation of Policy Effects

As described in Section 1.2, the Airbnb share as well as the number of remodels, additions, and advance information requests, increased following the Airbnb reform. Moreover, this spike is more pronounced for touristy civil parishes. To measure the impact of this policy intervention, we adopt a DiD methodology. The focus of this approach is on the difference between touristy (treatment group) and other civil parishes (control) (first difference), and the change that took place following the Airbnb reform in 2014Q3 (second difference).

More specifically, our DiD exercise is to test whether touristy civil parishes showcase a higher increase in the number of permits than other civil parishes with Airbnb reform (that is, post-2014Q2). Thus, our DiD approach first compares the evolution of the average number of permit requests of touristy civil parishes against other parishes.

Our main specification includes interaction variables consisting of a treatment dummy (1 if the treatment is active in civil parish i; 0, otherwise) and a reform dummy (1 if a quarter-year is post-2014Q2) as follows:

$$Y_{it} = \alpha_1 \; Reform_t \times Treatment_i + \tau_t + \mu_i + \varepsilon_{it} \tag{1.2}$$

In this specification, we let Y_{it} denote the natural log of the number of permit requests, for civil parish i, in period t (a given quarter-year). The other terms follow the same terminology as equation (1.1) with the difference being the interaction $Reform_t \times Treatment_i$ which captures the average impact since 2014Q2. It does so by comparing Y_{it} during the pre-reform against the post-reform quarters, as well as the treatment group versus the control group. Once again, the time-fixed effects trace out the common time trend, while the civil-parish fixed effects control for the ex-ante sources of heterogeneity (omitted variables).

While the exogeneity of the Airbnb reform might be a plausible assumption, for a valid DiD specification we require two additional identifying assumptions: parallel trends or no differential pre-trends; and stable unit treatment value assumption (SUTVA).

Stable unit treatment value assumption (SUTVA). This requires that the response of a particular unit should not be affected by the assignment of treatment to the other units. However, in urban settings such as this one, where treatment areas may affect adjacent non-treatment areas through equilibrium effects **Baum-Snow and Ferreira, 2015** this might be difficult to achieve. To overcome this issue, we aggregate our data to the highest possible geographic unit, which in our case is the civil parish level. Furthermore, it should also be noted that within the context of our analyses, a violation of the SUTVA would decrease the treatment effect, as a strong treatment would increase the number of permit requests not only in its civil parish, but also in adjacent civil parishes, reducing the gap between the two (i.e., between treatment and control, or touristy and other parishes). Thus, in the case of violation of SUTVA, the estimated treatment effect generated by the number of permits becomes a lower bound and our results are a conservative estimate of the true effect.

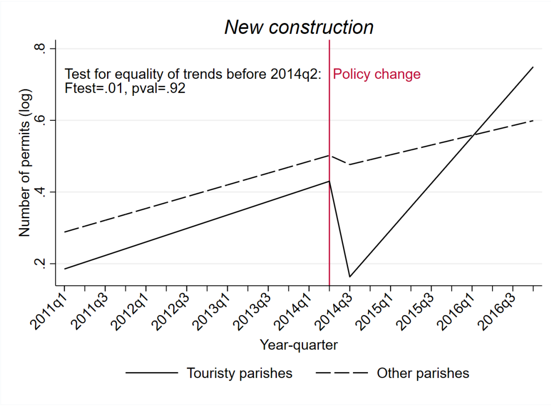
Parallel trends. A second important assumption is that of parallel trends in the dependent variable, before the policy change (pre-2014Q2), for both treatment and control groups. This assumption requires that the other civil parishes provide the appropriate counterfactual of the trend that the touristy civil parishes would have followed if they had not been treated. In Figure 1.7, we plot the values of our dependent variables smoothed by a linear trend for the touristy (treatment) and other parishes (control), before and after 2014Q2 (the last period without reform). Additionally, we perform and report in the same figures, a formal statistical test for equality of trends by estimating a model similar to our main specification in equation (1.2), using data from 2011Q1 to 2014Q2, except that the key variable of interest is an interaction between the linear time trend and the treatment group dummy instead. The test results fail to reject the equality in parallel linear trends in the pre-treatment data for the different types of permits, which is in line with the visual inspection of the trends presented in the figures.

Furthermore, we conducted an event study exercise to confirm that, prior to the Airbnb reform, the outcome variables in the treated and control parishes had parallel trends. The omitted quarter is the one right before the implementation of the policy (2014Q2). The following specification is used to conduct the event-study exercise:

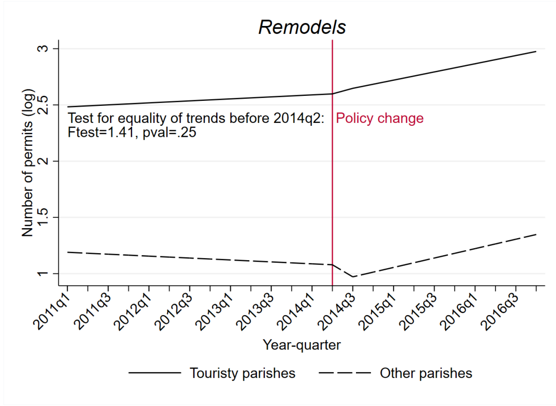
$$Y_{it} = \beta_0\ Treatment_i + \beta_{1it} Quarter_t \times Treatment_i + \tau_t + \mu_i + \varepsilon_{it} \qquad (1.3)$$

The symbology is exactly the same one used for equation (1.2), except for $Quarter_t$ which refers to the quarter dummies. We present the results in the Appendix – Figure A1 and plot the coefficients β_1 for the interaction $Quarter_t \times Treatment_i$, which mostly confirms that permit requests followed parallel trends before the reform.

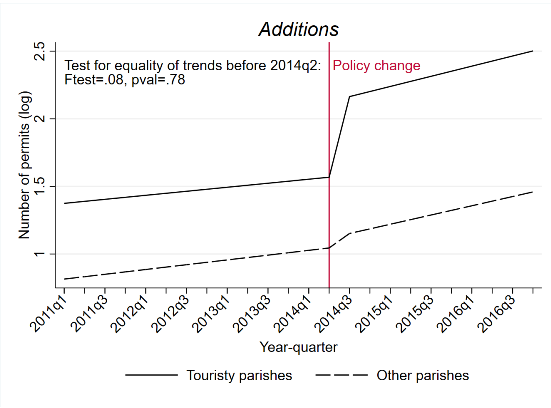
Figure 1.7: Estimated linear trend for treatment and control group



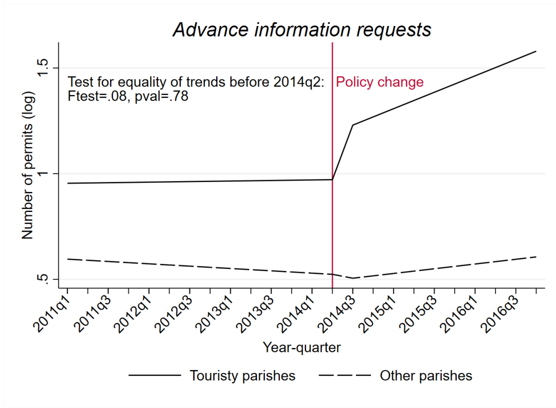
(a)



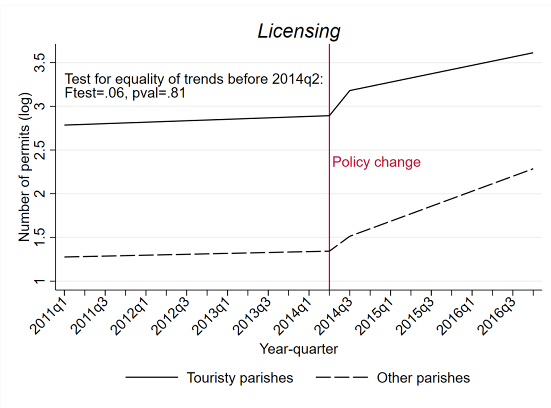
(b)



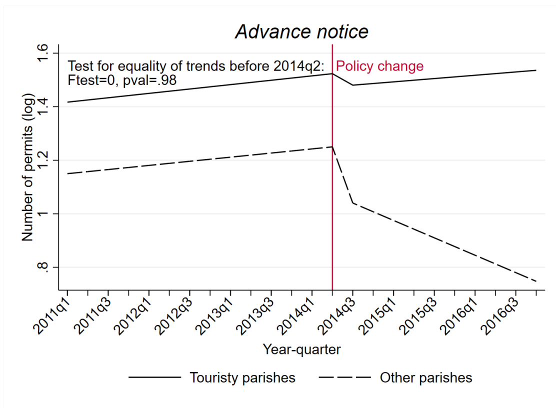
(c)



(d)



(e)



(f)

Note: Panels (a) - (c) refer to investment types (i.e., permit subtypes), while (d) - (f) represent permit types.

We also conduct a complementary specification following an IV approach. This allows us not only to validate the results from our main specification but also to obtain an estimate for the elasticity of permit requests with respect to Airbnb share. More concretely we predict, in the first stage, Airbnb share in the municipality (or civil parish) i and time t, as a function of the interaction between the treatment and reform dummies. This should account for the possible endogeneity between Airbnb share and permit requests and can be written as:

$$AbbShare_{it} = \beta_1 Reform_t \times \ Treatment_i + \tau_t + \mu_i + \varepsilon_{it} \tag{1.4}$$

In the second stage, we compute the log of permit requests as a function of the predicted Airbnb share and the same fixed effects used in the first stage:

$$Y_{it} = \alpha_2 \ \widehat{AbbShare}_{it} + \tau_t + \mu_i + \varepsilon_{it} \tag{1.5}$$

Relevance and exclusion restriction. The validity of $Reform_t \times \ Treatment_i$ as an instrument rests on its ability to satisfy the relevance and exclusion restrictions. The relevance assumption in our context translates into the touristy parishes being more affected by the reform than the other areas, in terms of an increase in the Airbnb share following the policy change. On Section 1.2 we illustrate exactly this. Also, the results of the formal test are presented in the results table (Table 1.5). The exclusion restriction assumption requires that the instrument, represented by the interaction between treatment and reform dummies in our study, should not directly influence permits, except through its impact on Airbnb share. Additionally, the instrument should not be correlated with other unobservable factors that influence the number of permits. Although it is not possible to directly test the exclusion restriction, we have confidence in its validity for the following reasons: Firstly, we did not observe any divergent pre-trends, indicating that the timing of regulations is not correlated with unobservable factors affecting permit requests. Secondly, the Airbnb reform was specifically targeted at short-term rentals. Lastly, we are unaware of any other land use or housing policies during this period that align with the Airbnb reform. Therefore, we have reason to believe that these highly specific regulations would primarily impact short-term rentals and not residential investment in any other way.

Finally, we conduct an analysis of the policy effects by the length of exposure by employing the following specification which follows the same terminology as equation (1.2), except for $Year_t$ which represents the years after the reform (i.e., 2014, 2015 and 2016):

$$Y_{it} = \alpha_3\ Reform_t \times Treatment_i \times Year_t + \tau_t + \mu_i + \varepsilon_{it} \tag{1.6}$$

1.4.3 Results

Table 1.3 presents our main results: the effects of the Airbnb reform on the different types of real estate investment. We find that, in comparison to other parishes, touristy parishes had, on average, an increase of 18.4% and 38.2%, and for remodels and additions, respectively, as a result of the Airbnb reform (i.e., after 2014Q2). Furthermore, in column (1), we also see that the effect for new construction is negligible and not statistically significant. These results are in line with the descriptive evidence presented in Section 1.2.

In Table 1.4 we show the results for the different permit types. After the 2014Q2 Airbnb reform, we note a 41.8% average increase in advance information requests relative to other parishes. The 34.1% average difference in advance notices merits cautious interpretation, particularly given the unanticipated decline in the control group post-reform, as highlighted in Figure 1.5. We suggest that without the policy, the treatment group could have experienced a similar decline. The policy might have not only bolstered the treatment group's outcomes but also shielded it from potential drops, underscoring its dual impact. Conversely, licensing appeared unaffected by

Table 1.3: DiD estimate of the effect of the policy on investment types

Dependent variable:	N. permit requests (log)		
Investment type:	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)
Overall treatment effect	0.060	0.184**	0.382***
	(0.087)	(0.075)	(0.105)
Quarter-year FE	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes
Observations	287	522	513
R-squared	0.621	0.936	0.884

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Average treatment effects of Airbnb reform on permit types and sub-types. The treatment group includes the civil parishes of Santo António, Santa Maria Maior, Misericórdia, Arroios, Estrela, and São Vicente. The control group comprises the remaining 18 parishes in the municipality of Lisbon.

the reform, as evidenced by a coefficient that was not statistically significant – and even negative.

Using our supplementary IV method, we determined that a 1pp increase in the Airbnb share due to the reform is associated with a 1.88% increase in remodels and a 3.82% rise in additions. Nonetheless, the results for new construction are not statistically significant and the coefficient is considerably smaller. When we examine permit types, our findings also align with the primary specification: with a 1pp uptick in Airbnb share there is a 4.26% surge in advance information requests and a 3.45% increase in advance notices. Again, licensing appears unaffected by the policy, even displaying a contrary trend. It is worth noting that our diagnostics show statistically significant first stage results across all permit types and subtypes.

The varying levels of Airbnb presence among civil parishes indicate that the impact of Airbnb is not uniform across all areas. This observation is demonstrated in Figure 1.8, where we plot the product of our coefficients in Table 1.5 with the Airbnb share of each civil parish at the end of our period of analysis (i.e., 2016Q4). While the effects are negligible for the less centrally located civil parishes, our results reveal significant local impacts for the touristy area. For instance, whereas Santa

Table 1.4: DiD estimate of the effect of the policy on permit types

Dependent variable:	N. permit requests (log)		
Permit type:	*Licensing*	*Advance notice*	*Advance information requests*
	(1)	(2)	(3)
Overall treatment effect	-0.148	0.341**	0.418***
	(0.090)	(0.143)	(0.130)
Quarter-year FE	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes
Observations	552	509	411
R-squared	0.951	0.851	0.707

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Average treatment effects of Airbnb reform on permit types and sub-types. The treatment group includes the civil parishes of Santo António, Santa Maria Maior, Misericórdia, Arroios, Estrela, and São Vicente. The control group comprises the remaining 18 parishes in the municipality of Lisbon.

Table 1.5: IV estimate of the effect of the policy on permit and investment types

Dependent variable:	N. permit requests (log)					
Permit type/ investment type:	*Licensing*	*Advance notice*	*Advance information requests*	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)	(4)	(5)	(6)
Airbnb share	-1.490	3.450***	4.259***	0.809	1.876*	3.824***
	(0.975)	(1.111)	(1.228)	(1.663)	(0.961)	(1.065)
Quarter-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	552	509	411	287	522	513
IV first stage (Fstat)	505	462	364	240	475	466
IV first stage (pval)	0.000	0.000	0.000	0.000	0.000	0.000
Endogeneity (pval)	0.775	0.038	0.030	0.996	0.173	0.010

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Instrument given by the interaction between the dummies for touristy parish and policy period.

Clara showcases an estimated 0.5% average increase in remodels as a result of the Airbnb reform, Santa Maria Maior displays a 74.9% surge.

Finally, we can see in Table 1.6 the policy's impact on remodels and additions peaked in 2015. In contrast, new construction does not display any consistent pattern, with most of the considered time horizon not being statistically significant. This supports prior findings that the reform did not influence new construction. The policy's effect on advance notices and information requests seems to commence in 2015 and remains more sustained. Meanwhile, licensing patterns remain undistinguished, mirroring the behavior of new construction. Note that because the reform was implemented in 2014Q3 we are only capturing half of the 2014 as opposed to the full years of 2015 and 2016.

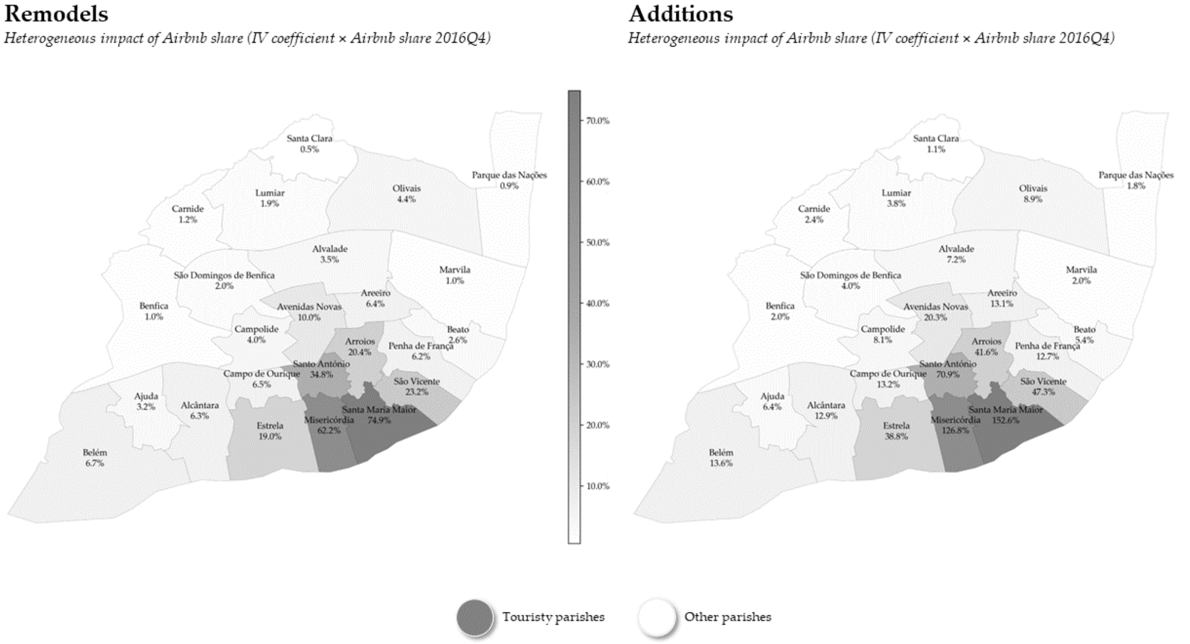
Table 1.6: DiD estimate of the effect of the policy on permit and investment types across time

Dependent variable:	N. permit requests (log)					
Permit type/ investment type:	*Licensing*	*Advance notice*	*Advance information requests*	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment x Reform x 2014	0.167	0.009	0.293	0.191	0.177*	0.306*
	(0.140)	(0.243)	(0.223)	(0.251)	(0.102)	(0.177)
Treatment x Reform x 2015	-0.112	0.506***	0.429***	-0.155	0.269**	0.629***
	(0.114)	(0.156)	(0.148)	(0.132)	(0.120)	(0.118)
Treatment x Reform x 2016	-0.193	0.457**	0.511***	0.268*	0.154	0.337***
	(0.122)	(0.170)	(0.130)	(0.135)	(0.102)	(0.104)
Quarter-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	552	509	411	287	522	513
R-squared	0.957	0.861	0.710	0.629	0.937	0.900

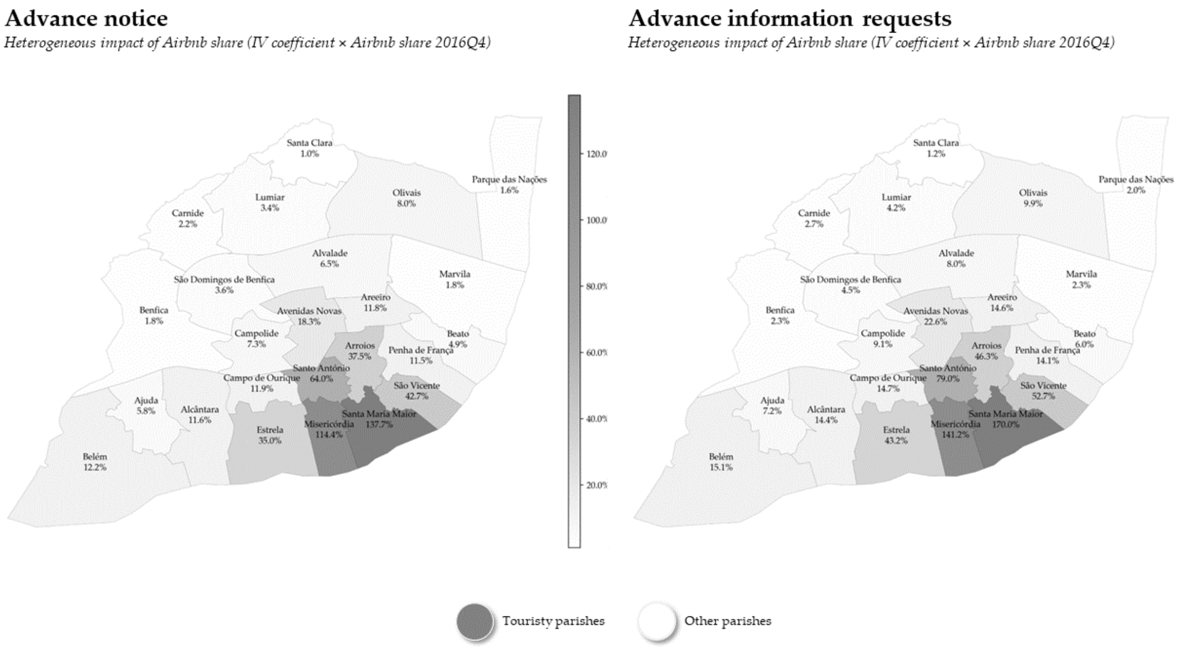
Note: Significance levels: ***p $<$ 0.01, **p $<$ 0.05, *p $<$ 0.1. Standard errors in parentheses clustered at the civil parish level. Average treatment effects of Airbnb reform on permit types and sub-types. The treatment group includes the civil parishes of Santo António, Santa Maria Maior, Misericórdia, Arroios, Estrela, and São Vicente. The control group comprises the remaining 18 parishes in the municipality of Lisbon.

All in all, our empirical results support that Airbnb reform stimulated the housing supply in touristy parishes, through the renewal of the local housing stock (in the form of remodels) and the amplification of existing properties (explained by additions). On the other hand, these touristy parishes coincide with areas where physical space and legislation constrain new construction. Thus, property owners are

Figure 1.8: Spatial distribution of policy effects (IV approach results)



(a)



(b)

Note: Panel (a) refers to investment types (i.e., permit subtypes), while (b) represents permit types. The remaining investment and permit types were not presented given that the resulting coefficients were not statistically significant.

left with the alternatives of expanding, whenever possible, or renewing the dwelling to increase capacity or price in the STR market. This also explains the absence of an effect of the Airbnb reform on the new construction. Moreover, we provide evidence that investors, seemingly driven by a desire to rapidly harness the advantages of the reform, displayed a strategic inclination towards swifter permitting processes. Our evidence underscores this, highlighting a marked shift in permit types with shorter implementation periods—specifically, advance notice and advance information requests—towards touristy civil parishes.

1.4.4 Robustness Tests

Two issues that might undermine our results are the following: (i) the policy intervention date; and (ii) treatment group composition. As robustness tests, we rerun our main specification (see equation (2)) using an alternative policy change and a different treatment group, which comprises the historical parishes of the municipality of Lisbon instead. This ensures that our results are not influenced by arbitrary selection of the policy reform date or by our specific definition of 'touristy' civil parishes.

Placebo Test

We conduct a placebo test by slicing our time horizon to consider only the period before the Airbnb reform (pre-2014Q3) and two alternative policy interventions, which took place in the same quarter (2012Q3). The first one is a policy aimed at liberalizing the rental market, while the second one is a reform that facilitates the visa attribution ("Golden visa") process to investors in the Portuguese real estate market.[18] We rerun our DiD specification for the period 2011Q1 – 2014Q2, using as a placebo the policy change of 2012Q2 (treatment given post 2012Q2) and present the results in Table 1.7. In comparison with the results obtained in Table 1.3 and 1.4,

[18] Market liberalization by *Law 31/2012* of 14$^{\text{th}}$ August 2012 and "Golden Visa" by *Law 29/2012* of 9$^{\text{th}}$ August 2012.

the coefficients shrink and are not statistically significant, suggesting the effects vanish for this period and alternative policy intervention in 2014Q3. This provides support for the results in our main specification being caused by the Airbnb reform.

Table 1.7: Placebo test using the "Golden visa" policy (2012Q2)

Dependent variable:	N. permit requests (log)					
Permit type/ investment type:	*Licensing*	*Advance notice*	*Advance information requests*	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)	(4)	(5)	(6)
Overall treatment effect	0.003	0.087	0.015	0.145	0.134	0.047
	(0.136)	(0.149)	(0.257)	(0.160)	(0.140)	(0.119)
Quarter-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	312	309	226	133	301	289
R-squared	0.942	0.856	0.644	0.603	0.941	0.835

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Average treatment effects of Airbnb reform on permit types and sub-types. The treatment group includes the civil parishes of Santo António, Santa Maria Maior, Misericórdia, Arroios, Estrela, and São Vicente. The control group is composed of the remaining 18 parishes in the municipality of Lisbon. Placebo test for 2012Q2 as an alternative policy change date and data constrained to the pre-Airbnb reform period (2011Q1 - 2014Q2).

Historical Parishes

Instead of using the touristy civil parishes as the treatment group, we use "historical" ones, which include historically protected areas, as defined by the municipality of Lisbon, to test the robustness of our results.[19] It is important to note that there is some overlap between the touristy and historical parishes, with 4 civil parishes being part of both groups.[20] We rerun our main specification for our entire time horizon (2011Q1 – 2016Q4), using the main policy change date (2014Q3), and obtain the results in Table 1.8.

Once again and similarly to the placebo test conducted in the previous subsection, in comparison with Tables 1.3 and 1.4, the coefficients reduce significantly, and the results are not statistically significant. This is an important result given that

[19]Historical parishes include the following: Santa Maria Maior, Santa Clara, Misericórdia, Estrela, Lumiar, Olivais, and São Vicente. The control group comprises the remaining 17 parishes in the municipality of Lisbon.

[20]The civil parishes of Santa Maria Maior, Misericórdia, Estrela, and São Vicente are classified as both *historical* and *touristy*.

Table 1.8: DiD estimate of the effect of the policy on "historical" parishes

Dependent variable:	N. permit requests (log)					
Permit type/ investment type:	*Licensing*	*Advance notice*	*Advance information requests*	*New construction*	*Remodels*	*Additions*
	(1)	(2)	(3)	(4)	(5)	(6)
Overall treatment effect	-0.027	0.146	0.046	0.064	0.029	0.165
	(0.136)	(0.181)	(0.151)	(0.107)	(0.117)	(0.143)
Quarter-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Civil parish FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	552	509	411	287	522	513
R-squared	0.950	0.849	0.699	0.621	0.936	0.881

Note: Significance levels: ***p < 0.01, **p < 0.05, *p < 0.1. Standard errors in parentheses clustered at the civil parish level. Average treatment effects of Airbnb reform on permit types and sub-types. The treatment group includes the civil parishes of Santa Maria Maior, Santa Clara, Misericórdia, Estrela, Lumiar, Olivais, and São Vicente. These correspond to the historical parishes which include historically protected areas, as defined by the Municipality of Lisbon. The control group is composed of the remaining 17 parishes in the municipality of Lisbon.

requalification of the housing stock found in our main results could have been due to the recovery of the derelict buildings int the historical centers. However, what we observe is that indeed requalification of the housing stocks occurs in parishes where STR presence is strong. When we consider instead other historical parishes with less STR presence, the results get diluted thus confirming the role of STR on stimulating housing supply.

1.5 Conclusion

In this chapter, we measure the effect of STR, more specifically of Airbnb, on housing stock. We use a DiD approach to show that between 2014Q3 and 2016Q4, following a policy intervention that facilitated the entry into the STR market, there was an increase in the number of real estate permits issued, for the areas with the strongest presence of Airbnb share (i.e., touristy civil parishes).

Our estimates suggest that the STR regulation increased remodels and additions by about 18.4% and 38.2%, respectively, for the touristy parishes when compared with the non-touristy ones. Notably, there was no observed impact on new construction. Furthermore, we have demonstrated that investors, likely motivated to quickly capitalize on the reform's benefits, strategically leaned towards expedited permitting processes. This is evident from the pronounced shift in permit types, particularly advance notice and advance information requests, favoring touristy civil parishes by 34.1% and 41.8% respectively.

Our supplementary IV method offers deeper insights into these dynamics. We found that a 1pp increase in the Airbnb share due to the reform is linked to a 1.88% increase in remodels and a 3.82% uptick in additions. As for permit types, we observed a 4.26% rise in advance information requests and a 3.45% surge in advance notices for every 1pp increase in Airbnb concentration.

Overall, our results indicate the Airbnb reform stimulated housing supply, creating incentives for improving the housing stock, promoting property remodels and additions, as well as increasing the interest in real estate investment – as given by the increase in the advance information requests – in touristy parishes. Moreover, investors preferred faster permit types (advance notice) to longer ones (licensing). However, this impact did not necessarily translate into new housing stock, as these civil parishes are characterized by physical or legal constraints for new construction. Our robustness tests, which alter both the composition of the treatment group and the policy date, corroborate that our results are indeed attributable to the Airbnb reform and are not arbitrary.

The documented change in the housing stock can subsequently drive up housing prices, not just due to a reallocation from long-term to short-term rentals, but also from the introduction of higher-end urban amenities, services, and enhanced infrastructure. As policymakers assess the implications of Airbnb's concentration on housing prices, it is vital to recognize the multifaceted nature of these price dynamics. While this chapter underscores the link between housing stock changes and rising Airbnb share, quantifying its direct impact on housing prices (both in terms of house transactions and rents) remains an avenue for future research.

1.6 Appendix

Table A1: Number of monuments and hotels

Civil Parish	**Touristy**	**Monuments**	**Hotels**
Ajuda	-	2	0
Alcântara	-	2	1
Alvalade	-	0	3
Areeiro	-	1	5
Arroios	Yes	4	18
Avenidas Novas	-	1	24
Belém	-	7	1
Campolide	-	0	5
Carnide	-	2	0
Estrela	Yes	4	1
Lumiar	-	1	0
Marvila	-	1	0
Misericórdia	Yes	5	2
Parque das Nações	-	0	5
Penha de França	-	1	0
Santa Maria Maior	Yes	21	19
Santo António	Yes	5	34
São Domingos de Benfica	-	3	1
São Vicente	Yes	3	1
TOTAL		58	119

Note: Touristy civil parishes correspond to the ones with the largest Airbnb share (in 2014Q2, before the policy change), given by the ratio of cumulative Airbnb listings to the total number of dwellings. Number of hotels that were open before 2014 (i.e. the year of the policy implementation).

Table A2: Issued building permits

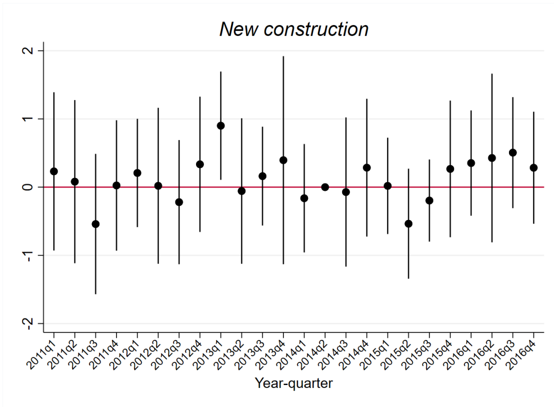
Year	2011	2012	2013	2014	2015	2016
New construction						
Number of issued building permits	46	27	17	22	28	32
Share of issued building permits for residential use	70%	85%	82%	73%	57%	91%
Remodels and additions						
Number of issued building permits	452	659	146	511	366	314
Share of issued building permits for residential use	74%	73%	79%	71%	70%	86%
TOTAL permits issued						
Number of issued building permits	498	686	163	533	394	346
Share of issued building permits for residential use	74%	73%	80%	71%	69%	87%

Table A3: Treatment and control groups' comparison

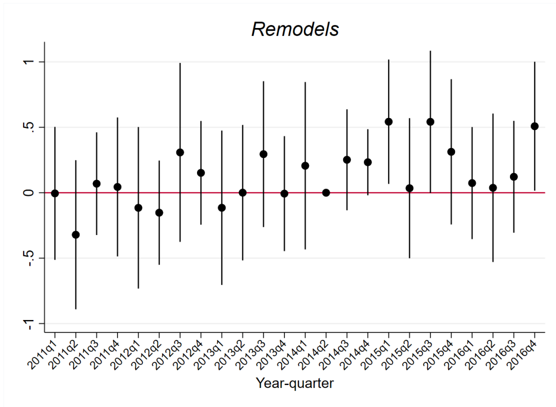
Variable	Control	Treatment
Number of parishes	18	6
Airbnb share	0.12%	1.70%
Derelict buildings (%)	6.44%	10.61%
House density (N/ km2)	3,916	8,255
Population density (N/ km2)	6,738	11,096
Dwelling age (years)	54.26	80.07
Number of floors per building	3.99	3.70
Number of dwellings	13,985	12,509

Note: Averages for treated and control civil parishes for the pre-policy period (before 2014Q3).

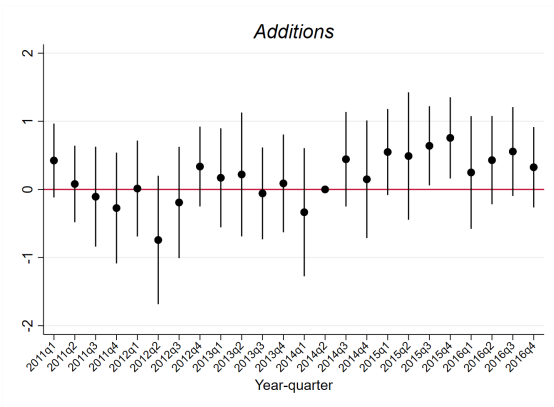
Figure A1: Event study



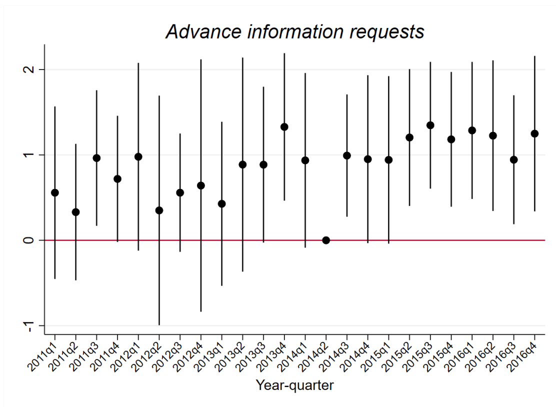
(a)



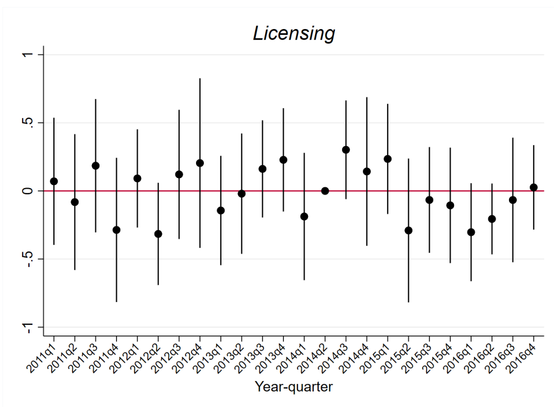
(b)



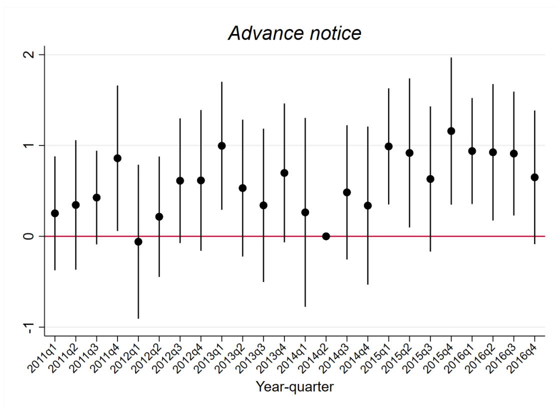
(c)



(d)



(e)



(f)

Note: Effects in comparison to 2014Q2 (period before policy shock). The treatment group includes the civil parishes of Santo António, Santa Maria Maior, Misericórdia, Arroios, Estrela, and São Vicente. The control group comprises the remaining 18 parishes in the municipality of Lisbon. The regression includes civil parish and quarter-fixed effects. 95% confidence intervals. Panels (a) - (c) refer to investment types (i.e., permit subtypes), while (d) - (f) represent permit types.

Chapter 2

Enhancing Data for Real Estate Policy: The Case of Portugal

Bernardo Forbes Costa, Fábio dos Santos Cardoso, Leid Zejnilovic

2.1 Introduction

The real estate sector holds substantial importance in modern economies. It represented 12.4% of the Gross Domestic Product (GDP) in the US in 2022.[1] In the same year, the EU and Portugal accounted for 10.3% and 12.1% of Gross Value Added (GVA), respectively.[2] It has a multidimensional role in driving economic activity. Housing markets are essential for financial stability, with housing price bubbles severely impacting the economy (Òscar Jordà et al., 2015). Furthermore, housing represents a significant portion of household wealth and is key in driving private consumption (Mian et al., 2013).

More importantly, real estate is probably one of the main topics in policymakers' current agendas. Rohe (2017), Wetzstein (2017), and Galster and Lee (2021) discuss how the housing affordability crisis we are currently facing has broad socio-economic implications. The authors suggest it has increased homelessness, reduced homeownership rates, escalating rental costs, increased economic strain for households, and even worsened health outcomes. They identify as causes factors like rapid urbanization leading to increased demand for housing, income disparities where wage growth does not keep pace with housing costs, and, most importantly, policy inadequacies that fail to address these issues effectively. To effectively address this crisis, bridging the 'policy-outcome gap' Wetzstein (2017) is crucial. By grounding policymaking in empirical evidence – that is, policies that are 'evidence-based' – and making them more effective, policymakers can develop strategies more aligned with the concrete challenges of housing affordability (Wetzstein, 2017; Rohe, 2017).

According to Baron (2018), evidence-based policy (EBP) is an approach to policymaking that relies on data analysis, empirical evidence, and research to inform decisions and interventions. The author outlines that EBP policy is associated

[1] U.S. Bureau of Economic Analysis, Value Added by Industry: Finance, Insurance, Real Estate, Rental, and Leasing: Real Estate and Rental and Leasing as a Percentage of GDP [VAPGDPRL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/VAPGDPRL.

[2] Eurostat data.

with several benefits, encompassing economic and social outcomes across various domains. Economically, it can result in efficient resource allocation, promote cost-effective interventions, and maximize return on investments. It focuses public and private resources on programs and practices with demonstrated effectiveness, optimizing the utilization of public funds and contributing to overall economic growth. Moreover, as discussed by Hart and Yohannes (2019), and Arinder et al. (2016), by prioritizing interventions informed by science, policies can better contribute to positive social indicators, such as improved health, higher educational attainment, reduced poverty, and enhanced social cohesion. Furthermore, the authors propose that EBP promotes the availability and sharing of research findings to increase transparency in accessing and communicating the evidence for policymaking processes. This ultimately leads to a more transparent and well-informed policy culture and constituency. Hart and Yohannes (2019) compilation of EBP cases brings to light compelling examples of evidence-informed policymaking's efficacy.

EBP initiatives in real estate have gained momentum globally, driven by the housing affordability crisis and a shift towards data-driven approaches in policy development. The #Housing2030 initiative is a key global endeavor in this domain. It harnesses the expertise of over 100 researchers, policymakers, and housing providers from the United Nations Economic Commission for Europe (UNECE) region and beyond. Collaborating with entities like Housing Europe and UN-Habitat, this initiative aims to define and implement effective policies that enhance housing affordability and sustainability, showcasing their real-world applications across diverse contexts. In the US, the Department of Housing and Urban Development (HUD) has a rich history of incorporating research into its policies, with its focused efforts on EBP intensifying in recent decades. In Europe, Norway's Housing Lab and the UK's Collaborative Centre for Housing Evidence (CaCHE) represent significant strides in evidence-based housing initiatives. Similar trends are evident in Australia and Canada, with the Australian Housing and Urban Research Institute (AHURI) and the Canadian Housing Evidence Collaborative (CHEC), respectively. Despite differing geographical and cultural contexts, these initiatives are united in their commitment to leveraging data and research for crafting effective real estate policies. This global trend underscores the growing recognition of the value of evidence-based approaches in addressing housing challenges.

Even though it can be argued that academics and policymakers view and use 'evidence' differently, research is the first cornerstone for EBP (Arinder et al., 2016). According to Hjort et al. (2021), it can shape policymakers' beliefs about the effectiveness of policies and prompt the adoption of evidence-informed ones. The authors suggest that informing policymakers about research on effective policies increases the likelihood of policy implementation. Furthermore, they point out that utilizing specific types of studies and methods, such as larger-sample studies, has been shown to influence how political leaders perceive research findings and lead to updates in their beliefs. The second and most important cornerstone of EBP is data. The role of data is significant for both research and EBP directly. Data serves as the foundation for evidence generation in research, enabling the analysis and evaluation of phenomena to produce meaningful insights (Guardia et al., 2020). Ooijen et al. (2019) assert that high-quality and specific data contribute to the credibility and robustness of research findings, ultimately enhancing the reliability and trustworthiness of research outcomes. According to Ooijen et al. (2019), in the public sector, this data is not only an asset to policymaking but also to service delivery, organizational management, and innovation. For policy formulation, using relevant and actionable information obtained typically from specific data is important to assess the potential impacts of various policy alternatives (World Bank, 2021). Therefore, data holds a dual role, serving as the backbone of research and as a lever in shaping EBP.

However, it is not only data but data access that is crucial. According to Berends et al. (2020), open data refers to data that is open in terms of access, redistribution, reuse, absence of technological restriction, attribution, integrity, and no discrimination. This means the data is freely available for anyone to access, use, and share. Public entities often produce or commission this data as part of open government initiatives, commonly called Open Government Data (OGD). Still, it also extends to privately held data (i.e., open industry data) and can be of general public interest for various purposes, such as commercial, research, or innovation.[3] The importance of open data for EBP cannot be overstated, though. This is illustrated by the fact that the OPEN Government Data Act is an integral component of the US Foundations for Evidence-Based Policymaking Act passed in 2019. It mandates federal data's openness and machine-readability to enhance governmental accountability and foster data-driven economic growth. Concurrently, the European Union (EU) passed the Directive (2019/1024) on Open Data and the Reuse of Public Sector Information. This directive focused on making high-value datasets freely available and harmonizing data access and formats across EU Member States to promote and bolster economic innovation and transparency. Both initiatives reflect a growing global consensus on the economic and democratic value of open data, significantly shaping evidence-based policymaking.

Nagaraj et al. (2020) highlights that improved data access (i.e., data is made openly available) has been shown to increase scientific research's quantity, quality, and diversity. The authors note that when barriers to data access are lowered, it is more likely to be utilized by scientists, leading to a greater quantity and quality of scientific output. Nagaraj and Tranchero (2023) discuss the multifaceted benefits of improved data access on scientific research. They state that improved data access leads to a higher diffusion of data, which in turn facilitates empirical researchers in publishing more articles in prestigious outlets and increasing citation-weighted publications. Furthermore, they emphasize that improved data access democratizes science by allowing authors with fewer financial resources to participate in the scien-

[3] A formal definition of open data can also be found at https://opendefinition.org/.

tific process, thereby increasing the diversity of scientific research. They argue this has the potential to diversify and broaden the scientific discourse. Most importantly, they conclude that data access not only boosts scientific progress but also significantly contributes to promoting EBP by increasing the consumption of research by policymakers, further facilitated by the improved scientific quality resulting from better data access.

As noted, open data provides a resource for pushing research and allowing for more informed and impactful policy decisions. However, detailed, specific, and up-to-date data on real estate has historically been difficult to obtain. Access to detailed information about home sales is limited in many countries and often comes with significant delays. Additionally, these datasets may have limitations related to their geographic or time-related scope and their informational accuracy (Loberto et al., 2021). Furthermore, open data in the real estate sector has received little attention, overshadowed by the more prominent focus on smart cities and broader urban economic concerns.[4] Nevertheless, some exceptions are worth mentioning that indicate a trend toward increased transparency and accessibility in real estate data. These are presented in Table 2.1. The increased availability of these databases marks a significant shift from the traditional lack of transparency in real estate data towards an open data culture.

Having established the importance of data as a lever for fostering EBP, in this paper we discuss how we can obtain more and better real estate data (i.e., enhance real estate data) so that we can close the ‘policy-outcome gap’ to address housing affordability and other complex real estate policy challenges. The remainder of the paper is then structured as follows. Section 2.2 defines real estate data, listing the wide range of currently available types. Section 2.3 discusses four distinct, but complementary, avenues for real estate data producers, in particular public institutions, to enhance real estate data. These include the adoption of big data, the

[4] See for instance Liu et al. (2015) and Adje et al. (2023).

TABLE 2.1: Real Estate Open Data Sources by Country

Country	Institution	Data Source	Details
UK	HM Land Registry	HM Land Registry Open Data	Property transactions with prices, dates, locations, property types, and additional transaction details.
Singapore	Housing & Development Board	HDB Resale Transactions	Property transactions with prices, dates, locations, property types, and additional transaction details.
Germany	RWI - Leibniz Institute for Economic Research	RWI-GEO-RED	Properties listed on online platforms, with details about prices, dates, locations, property types, and additional listing information.
Bosnia and Herzegovina	Federal Administration	State Property Register	Property ownership and transactions, with an interactive mapping tool.
Belgium	Association of Flemish Provinces	Belgium's Provinces in Figures	Housing and vacant real estate statistics across provinces.

use of Artificial Intelligence (AI) and blockchain technologies, and the undertaking of data linkage. We then apply these ideas to the case of Portugal in Section 2.4, examining the state of EBP and open data, analyzing the main institutions within the real estate sector, mapping the current real estate data landscape, and discussing the challenges with data production. Subsection 2.4.3 puts forward propositions, based on the insights from Section 2.3, to improve real estate data in the country. Section 2.5 offers remarks about our discussion and Section 2.6 concludes.

2.2 Real Estate Data

This paper primarily focuses on real estate data. Unlike urban data, which covers a broad spectrum of city-related information, including infrastructure, demographics, and public services, real estate data zeroes in on specific property prices, characteristics, and market trends. Also, it is more encompassing than housing, including commercial, industrial, and agricultural properties, in addition to residential spaces and the several stages of the real estate lifespan, including finance, planning, regulation, design, construction, management, refurbishment, and demolition. Real estate data offers insights into market dynamics, investment considerations, and socio-economic outcomes, which are relevant for research and policy formulation. For example, policymakers may need to implement different strategies based on whether soaring housing prices are driven by excessive demand or insufficient supply. This requires timely information on both demand and supply. Furthermore, this information must be accessible at the national level and a more granular local level. For instance, it has been shown that housing bubbles can emerge in local markets even when there are no indications of imbalances at the national level (Landvoigt et al., 2015). Moreover, access to granular data on house prices is crucial for understanding local phenomena, like neighborhood effects, which significantly impact socio-economic outcomes for residents (Chetty et al., 2016). Real estate data has broader applications, namely in environmental economics. Housing prices are a valuable indicator of consumer responses and preferences regarding environmental improvements such as enhanced insulation, renewable energy installations, and other green initiatives (Bauer et al., 2017).

The recent surge in the availability of diverse data sources has significantly enhanced the measurement and analysis of real estate dynamics. Drawing from the work of Weinberg (2014, 2015), Wei et al. (2022), and Saiz and Salazar-Miranda (2023), we provide a list of real estate traditional and big data (i.e., non-traditional) in Table 2.2. This list identifies optimal real estate data types that could be available to policymakers and researchers. Should any of these data types be unavailable, their absence becomes readily apparent, thus facilitating a critical discourse on the necessity and feasibility of their collection to enhance data availability. While we do not delve into detailed discussions of the studies employing this data or exploring additional data sources for urban and city planning, further insights can be found in the works of Wei et al. (2022) and Saiz and Salazar-Miranda (2023).

Table 2.2: Real Estate Data Types

Type	Description	Information
Financial and Mortgage Data	Offers insights into mortgage loans, rates, and financial products.	Data on mortgage origination, terms, rates, foreclosures, and defaults.
Credit Card Data	Reveals economic activity and spending trends through transactional data.	Indicators of neighborhood appeal based on consumer spending patterns.
Appraisals	Professional property valuations for sales, financing, and insurance.	Appraisal reports, valuation methods, and market analyses.
Historical Data	Insights from archived records on property values and development.	Long time series of transaction prices, land use changes, and real estate returns.
Prices and Affordability Indices	Changes in housing prices, volumes, and construction costs analysis.	Affordability indices, pricing data, transaction volumes, and construction cost indices.
Internet Search Behavior	Online search patterns to forecast real estate market demand.	Keyword search trends.
Online Advertisements	Market supply, demand, and valuation through online advertisement and other market platforms.	Listings of prices, features, and asking vs. selling prices comparison.
Textual Repositories	Housing market trends and policies from unstructured text data.	Analysis from news articles, internet posts, forums, and public records.
Multiple Listing Services	Comprehensive property databases for facilitating broker information exchange.	Property characteristics, historical listings, asking prices, broker details.
Street-view Imagery and Online Pictures	Visual data assessment of property conditions and neighborhood appeal.	Architectural style, materials, condition, size, and aesthetics.
GPS and CDRs Data	Commuting routes and accessibility analysis via GPS and CDRs.	Commuting patterns and transit node accessibility indexes.
Sensor Data	Environmental sensor data.	Air pollution, noise levels, and other environmental factors that can influence real estate preferences.
Volumetric Data and Street Shapes	Building morphologies and street layouts impact on property value.	Building heights, density, massing, and street layouts.
Governmental Data	Comprehensive housing information from census data and surveys.	Housing vacancies, tenure, characteristics, new construction, loans, demographics.
Construction and Building Permits	New construction activity tracking through permits and stages.	Units authorized, construction phases, building types, geographic distribution.
Housing Programs	Data on affordable and rural housing initiatives for underserved communities.	Loan programs, grants, subsidies, tax credits, affordability measures.
Geographic Information Maps	Spatial mapping for land dynamics, desirability, and real estate pricing.	Raster data, GIS polygon data, Points Of Interest (POIs), and GIS Polylines.
Public Registries	Property and land ownership, transactions, and historical data records.	Ownership documentation, transaction histories, and legal descriptions.
Brokers' Records	Real estate transaction data, reflecting demand and supply.	Listings and sales information, market analyses, broker insights.

Some data types are inherently more prone to being open than others due to the nature of their collection, their intended use, and the entities that manage them. For example, 'Governmental Data' and 'Geographic Information Maps' are typically openly available because public institutions often collect them with the mandate to make such information accessible to the public for transparency and to encourage civic engagement. On the other hand, data types like 'Credit Card Data' and 'Brokers' Records' are commonly 'closed', as they contain sensitive personal information or proprietary business information, making them subject to privacy laws and competitive considerations. Additionally, the level of aggregation or the granularity of the data—such as individual versus municipality—might influence the ability of data to be made openly available. Protecting the confidentiality of private information must be paramount, yet we should recognize that the private sector is already extensively utilizing sensitive data. Tene and Polonetsky (2011) argue that constraining access solely to the domains of research and policymaking could inadvertently neglect the societal benefits derivable from comprehensive data utilization. Thus, it is necessary to weigh both privacy concerns and the wider benefits to society when considering data openness.

2.3 Enhancing Real Estate Data

We have established that to tackle current real estate challenges, we need EBP, which in turn requires real estate data, namely open data. In other words, we need to enhance data to serve policies that are grounded in accurate, timely, and comprehensive information. The question then becomes: how can we achieve this? One of the alternatives and the one we pursue in this paper is to enhance the capabilities of institutions that are involved in real estate data generation and or distribution. This section delves into four possible avenues to do so: Firstly, we address harnessing big data to augment data availability and diversity. Secondly, we investigate the application of AI techniques for information extraction, classification, and browsing. Thirdly, we explore the utilization of blockchain technology to establish data processing and sharing infrastructures that inherently prioritize transparency and privacy. Lastly, we discuss data linkage and its implications.

2.3.1 Big Data

Big data can serve as a tool for modern policymaking, offering opportunities for enhancing policies' precision, responsiveness, and effectiveness in the real estate sector and beyond. As highlighted by Leoni et al. (2023), integrating non-traditional data types provides a multidimensional picture that traditional ones alone often cannot capture. The author suggests these novel data types, like 'Internet Search Behavior' and 'Online Advertisements,' can offer real-time insights into the complexity and interconnectivity of factors influencing real estate market dynamics, help predict market movements, and craft policies that are more closely aligned with current societal needs and behaviors.

Some examples might help illustrate how big data can be useful in real estate. Loberto et al. (2021) tap into data from 'Online Advertisements' to fill the informational voids left by traditional data sources, offering real-time insights into housing demand, supply, and pricing—information crucial for policymakers. Similarly, Hu et al. (2018) monitor fine-scale housing rental prices through online housing rental websites. Law et al. (2019) illustrates how the analysis of street-view imagery ('Street-view Imagery and Online Pictures'), alongside traditional housing attributes, unveils intangible real estate qualities that influence property values. Cao et al. (2021) employ text mining, using 'Textual Repositories', to develop indicators that gauge public opinions on real estate market policies. van Dijk and Francke (2018) explore 'Internet Search Behavior' data to examine housing market dynamics. These examples underscore the capacity of big data to unlock insights into market trends and policy impacts that might otherwise not be achieved with traditional data types, showcasing its value for evidence-based policymaking.

The argument here is that augmenting real data through big data utilization, or, put differently, broadening the array of real estate data types for empirical evidence generation, is of value. Initially, the list of real estate data types provided in Table 2.2 can be used to gauge the data needs of researchers and policymakers. Subsequentially, several guiding questions, adapted from Delisle and Grissom (2020), can help guide the decision to collect the data by relevant public institutions:

- What additional data types would we consider collecting if resources permitted?
- Why prioritize these data types over others?
- Can this data address the pressing challenges of our times?
- Will it remain relevant for addressing future questions?
- What trade-offs are associated with these data types?

Should the answers to these inquiries suggest that further data acquisition appears beneficial, a thorough cost-benefit analysis of data collection and management should follow. This approach ensures that the expansion of real estate data through big data enhances the depth and breadth of analysis and aligns with strategic priorities and resource constraints.

2.3.2 Artificial Intelligence

AI can enhance real estate data through data validation and verification, advanced information extraction, and document classification. By leveraging Machine Learning (ML) and Natural Language Processing (NLP) algorithms, it becomes possible to extract key information from diverse real estate document types, such as tenant details, lease expiration dates, and financial terms, streamlining the data entry process and reducing errors. Furthermore, AI can then use this information and automate the classification of documents into categories, such as leases, contracts, and maintenance records, enabling faster and more accurate organization of data (Bodenbender et al., 2019). AI can also improve data validation and verification by using feature learning, anomaly detection, and association rules, ensuring data reliability. Moreover, ML algorithms can cluster data automatically, streamlining the search for information and metadata (Treleaven et al., 2021). This improves data management and contributes to better empirical evidence by providing comprehensive, reliable, accurately classified, and easily accessible real estate data.

Cantador et al. (2021) underscore the significance of AI-driven chatbots in democratizing access to open data. This is important for public entities involved in data distribution, such as national statistics institutes. By employing NLP, these tools facilitate a conversational interface that simplifies user interaction with datasets, making the search and exploration process more intuitive and efficient. This not only broadens the accessibility of open data to a wider audience without technical expertise but also enriches the user experience, fostering data use. Building on this foundation, the advent of ChatGPT, as detailed by Alexopoulos et al. (2023), marks a leap forward in the capabilities of conversational interfaces for open data. ChatGPT, by offering more advanced, accessible, and engaging interactions, allows users to search, explore, process, and derive conclusions from open datasets, offering personalized recommendations, data visualization insights, and suggestions for future research areas. This improves the user experience and enhances the utility and impact of open data. ChatGPT also offers significant advantages to open data publishers, primarily government institutions, by providing insights into different user groups' data needs and preferences. This capability allows for a more demand-driven approach to data publication, ensuring that the available data is of high relevance and value to the community.

Integrating AI in real estate open data initiatives is poised to address key challenges previously faced, such as improving data quality, enhancing user engagement, and ensuring the strategic dissemination of data. However, these technologies must be applied ethically, with a conscious effort to mitigate algorithmic biases and uphold the principles of fairness and transparency (Treleaven et al., 2021).

2.3.3 Blockchain

Saari et al. (2022) detail how blockchain technology introduces transparency, security, and efficiency through a decentralized, immutable ledger. According to the authors, this innovation enhances data integrity by ensuring that information, once recorded, is tamper-proof, which allows for combating fraud and corruption in the real estate sector and building a foundation of trust. They further explain that the decentralized nature of blockchain democratizes access to real estate data, overcoming the traditional challenges of siloed information and restricted access. It facilitates a unified database that is openly shared among stakeholders, thereby improving transparency and inclusivity. Additionally, the paper notes that blockchain's use of smart contracts allows for higher efficiency in data management by automating processes, reducing administrative burdens, and providing timely, accurate information.

The concrete applications of blockchain in real estate are wide-ranging. Saari et al. (2022) survey the literature and identify several use cases of blockchain technology in areas such as real estate development, investment, administration, maintenance, leasing and renting, land administration, and property transactions. Among these, the authors highlight land administration and property transactions as areas where blockchain can impact. For land administration, the authors note that blockchain provides a secure method for recording land titles and registries, ensuring these records are easily accessible and protected against unauthorized changes, which is important in geographies prone to manipulating land ownership records. According to them, blockchain enhances the transparency and integrity of land registries, thereby helping to reduce fraud. Regarding property transactions, the authors explain that blockchain technology streamlines the process by minimizing the need for intermediaries, enabling more direct interactions between parties through smart contracts. These contracts facilitate rental agreements, payments, and the resolution of disputes more cheaply and efficiently, making buying, selling, and transferring property rights much more straightforward. This means that real estate data becomes more comprehensive, reliable, accessible, and easier to share as open data.

Lemieux and Flores (2017) and Mcmurren et al. (2018) provide real-world examples of how blockchain is being used to improve real estate data, particularly for land administration and property transactions (i.e., 'Public Registries' data), in Brazil and Sweden, respectively. In Brazil, Ubitquity, a company that specializes in blockchain-based recording of titles and ownership transfers, formed a collaboration with local registry offices and a local research group. This initiative aimed to leverage blockchain technology to streamline land administration and property transaction processes in Rio Grande do Sul, supported by further academic support from the University of British Columbia. Similarly, ChromaWay a blockchain technology company known for developing a platform for decentralized applications, partnered with a leading Swedish multinational telecommunications company and a local consulting firm to work with the Swedish Land Registry to improve the efficiency of property transactions and land ownership in Sweden. By creating a private blockchain-based transaction system, the project aimed to reduce the redundancy in data verification, showcasing the potential of blockchain to make real estate transactions more efficient and secure. Similarly in the UK, the HM Land Registry collaborated with ConsenSys Codefi to investigate how blockchain and smart contract technology can enhance efficiency, straightforwardness, and clarity in the real estate sector. These initiatives demonstrate the potential and feasibility of applying blockchain technology in the real estate sector. They also underscore the importance of collaboration among diverse stakeholders to ensure successful implementation, particularly with local registries.

Open data, particularly OGD, is typically centrally stored by one or few institutions. Janssen and van den Hoven (2015) notes the efficiency of such structures but also warns of privacy risks like data breaches and unauthorized use. This motivates the need for decentralized, transparent, and reliable information systems designed with privacy in mind, preventing the manipulation of critical information. This does not imply making all data accessible but guarantees that relevant information remains openly available. Privacy protection should be prioritized through access controls and data compartmentalization to maintain privacy standards and limit access to necessary data only, with strict sharing guidelines. As proposed by Joseph (2019), blockchain technology offers a solution to these open data challenges, enhancing its integrity, accessibility, and transparency. Firstly, data becomes immutable and verifiable, fostering trust among users. Secondly, blockchain enables a decentralized storage system, ensuring that data is secure and resistant to tampering and censorship. Thirdly, its inherent auditability feature promotes transparency, allowing users to track data provenance and modifications over time. Fourthly, implementing smart contracts on blockchain platforms can automate data sharing and access protocols, reducing administrative overhead and improving efficiency. Finally, blockchain facilitates the creation of a transparent ecosystem where stakeholders can participate and contribute to the governance of open data, ensuring that the data remains accessible, accurate, and relevant.

Despite the potential of blockchain-based solutions, challenges such as legal recognition of blockchain-generated titles, data security, and operational issues slow progress. To unlock the full potential of blockchain in the real estate industry, a unified effort from all stakeholders—registries and other public institutions, banks, buyers, and sellers—is then imperative (Saull et al., 2019).

As discussed by Tan (2021), the combination of big data, AI, and blockchain allows for creating a decentralized infrastructure capable of intelligent data analytics. According to the author, big data becomes more secure and valuable with blockchain technology, enabling verified and structured data for improved analysis. Tan explains that AI improves this by providing smart data administration and exploration. As the author describes, this integration makes data governance transparent, secure, and automated, improving data handling and governance and satisfying the data needs of several stakeholders, particularly policymakers.

2.3.4 Data Linkage

As discussed by Saiz and Salazar-Miranda (2023), some nations like Sweden and Denmark stand out in their efforts to provide extensive, unified datasets covering their entire populations. Given the challenge of implementing this type of centralized data ecosystem, other nations are alternatively linking and geocoding their datasets. In the US, for example, efforts to connect longitudinal information from diverse sources such as census data, social security records, and state unemployment databases have facilitated the creation of platforms like the Integrated Public Use Microdata Series (IPUMS), enabling the broad use of these data. Furthermore, projects that link historical census data have opened new paths for examining long-term trends, namely in the real estate domain. For instance, the HUD provides a linked dataset detailing housing needs, with a focus on the requirements of low- and moderate-income households, which supports state and local governments in formulating their Comprehensive Housing Affordability Strategy, components of eligibility for major grant programs. Utilizing these data enables grantees to conduct thorough assessments of housing and community development needs and market conditions, thereby facilitating informed, location-specific investment decisions (Weinberg, 2014). Thus, the linkage of several different datasets facilitates the revelation of insights that could not be derived from any single dataset alone.

However, this practice is not confined to public entities. As observed by Saiz (2020), private companies are increasingly collecting proprietary data and linking those datasets with publicly available ones. For instance, Zillow in the US collects data for past and current home supply from 'Multiple Listing Services' and links that to 'Governmental Data'. There are two main differences between public and private entities' efforts. First, while public bodies typically link traditional data types, private companies link traditional ones with big data. Secondly, whereas public organizations are linking datasets for research or policymaking purposes and perhaps making such data open, private institutions' primary goal is to monetize the data. Janssen and van den Hoven (2015) proposes the concept of 'Big and Open Linked Data (BOLD)', which refers to adopting big data, linking that data, and making it open. By doing this, public institutions can increase the value of their data while striving for transparency and creating competition for data oligopolies. This requires, however, a combination of data from different information silos, that is, data interoperability between institutions. Nevertheless, we must also consider privacy. A possible solution is to implement principles of transparency and privacy 'by design' in data ecosystems, as explored by Janssen and van den Hoven (2015). This approach balances the need for data openness with the imperative of safeguarding individual privacy by embedding these principles at the core of information systems architectures. As discussed before, blockchain could facilitate this by providing the appropriate data infrastructure that embeds these principles of transparency and privacy by default. Adopting BOLD sustained by blockchain infrastructure could counterbalance the dominance of data oligopolies by democratizing access to high-quality real estate data. Such an approach not only enhances the accessibility of public institutions' data for research and policymaking but also establishes a foundation for responsible data governance that respects individual rights and fosters a transparent, accountable public sector.

2.4 The Case of Portugal

For the case of Portugal, we start by examining the current state of evidence-based policy and open data and identifying the main institutions involved in the real estate sector. Following this, we examine the landscape of real estate data, highlighting the main data producers and outlining the issues encountered. Finally, drawing on the discussions so far, including augmenting data availability and diversity through big data, the use of AI and blockchain to improve real estate data processing and sharing, and promoting data linkage, we put forward propositions for public institutions in Portugal to improve the collection, availability, and use of real estate data.

2.4.1 Evidence-Based Policy and Open Data

According to Simoes et al. (2022), in Portugal, the integration of research into policymaking is paradoxical. On the one hand, policymakers recognize the scientific community's excellence. On the other, there is a lack of awareness of the importance and usefulness of empirical evidence. This contradiction hampers the potential of research to inform policy, as it is usually used for ad-hoc needs such as crisis response and the validation of pre-made decisions. Additionally, the authors suggest enhancing transparency in policymaking and promoting public engagement is crucial for the country. This requires more robust governance structures, emphasizing the importance of evidence-based policy and open data.

Portugal offers OGD through portals at both the national and municipal levels. The country's major cities, such as Porto and Lisbon, have established their dedicated portals for local data repositories, Portal de Dados do Porto and Lisboa Aberta, respectively, with the national platform being dados.gov.pt. However, within some of these platforms, the availability of real estate data is not as comprehensive as hoped. Annually, the EU issues the Open Data Maturity report (Page et al., 2023) to measure the progress of European countries in adopting and promoting open data. In 2023, Portugal showed an increase of 10 percentage points to 85% in its overall maturity score (which ranges from 0% to 100%), in line with the trend observed across the EU Member States – where the average maturity score increased from 79% in 2022 to 83% in 2023. This progress reflects the European effort to improve public sector information's availability and reuse, as guided by the already mentioned Directive (2019/1024), which in Portugal is encapsulated in Law No. 68/2021 (rectified by Rectification Declaration No. 31/2021), which formally recognized dados.gov.pt as the central catalog of open data in the country. Despite this progress, Portugal is positioned in the middle of the table, ranking 17th among 35 countries (including 27 EU members), indicating room for improvement.

In Portugal, the real estate sector is supported by several institutions presented in Table 2.3. These entities collectively ensure the sector's regulation, promotion, and development and are the producers of primary real estate data. While some of these institutions have launched impactful initiatives, like the Housing Portal, to assist with housing and broader real estate concerns, there is no dedicated real estate research division, contrasting with models from other countries such as the US's HUD, the UK's CaCHE, and Norway's Housing Lab. This absence points to a potential oversight in ensuring that policies and initiatives are underpinned by solid research, underscoring the urgency for more robust, evidence-based policymaking in Portugal's real estate sector.

2.4.2 Real-Estate Data Landscape

The sources of real estate data in Portugal are somewhat limited. A simple internet search will unveil to the reader some real estate data providers, but after careful analysis, she will realize that there are only a few primary data producers. A mapping of the several primary data providers is shown in Table 2.4.[5] Other initiatives,

[5] To collect the real estate data providers for Portugal, a search query was conducted on Google. The search terms used were 'dados imobiliário Portugal' ('real estate data Portugal' in English). Upon entering these terms into the Google search bar, the first three pages of the search results

Table 2.3: Institutions Supporting the Real Estate Sector in Portugal

Institution	Role
Housing and Urban Rehabilitation (IHRU)	Focuses on improving housing quality, urban renewal, and social housing.
Portuguese Association of Real Estate Developers and Investors (APPII)	Advocates for real estate promoters and investors.
Institute of Public Markets, Real Estate, and Construction (IMPIC)	Regulates public markets, real estate, and construction.
Institute of Registries and Notary (IRN)	Manages real estate ownership and transactions.
National Tourism Registry (RNT)	Oversees the regulation of short-term rentals.
Tax and Customs Authority (AT)	Handles rental contract agreements for taxation purposes.
National Energy Agency (ADENE)	Responsible for the certification and improvement of energy efficiency in real estate.
Municipality Halls	Regulate land use, issuing building permits, and ensuring compliance with local zoning and construction regulations.

like Lxhabidata or Pordata, are simply data aggregators from the sources presented.

The concerns with the Portuguese real estate data have been recently voiced by Rodrigues (2022). The report highlights several critical issues concerning the availability, accessibility, and granularity of data necessary to comprehensively understand and analyze the Portuguese real estate market dynamics. A primary challenge identified is the insufficiency of data and data restrictions, more pronounced than in other countries to analyze different drivers for house prices. This hinders a deeper understanding of the market dynamics. There is also a notable scarcity of spatial granularity. This situation is exacerbated by the fragmented nature of available databases and the restricted access to crucial attributes, such as house prices. Moreover, reconciling the availability of data at different geographic levels (ex. civil parish, municipality, national) is challenging. Lastly, the lack of central-

were thoroughly reviewed for relevant information. Readers should note that replicating the search may yield different Google results, influenced by the search algorithm's factors like location, search history, and device used.

Table 2.4: Real Estate Data Sources in Portugal

Data Source	Data Types	Open Data?
Statistics Portugal (INE) - Construction & Housing	Governmental Data, Prices and Affordability Indices, Financial and Mortgage Data	Yes
Bank of Portugal - Housing Prices	Prices and Affordability Indices	Yes
Idealista, Casa Sapo, Imovirtual	Online Advertisements	Yes
Confidencial Imobiliário	Multiple Listing Services	No
Portal de Dados do Porto, Lisboa Aberta, and others	Governmental Data	Yes
travelBI Open Data - Short-term Rentals	Governmental Data	Yes

ized registration mechanisms and Application Programmable Interfaces (APIs) for automatic data access from real estate online advertisement portals complicates the compilation of such data. These portals often limit data consultation mechanisms and lack features like property georeferencing and precise identification of transaction prices, rendering spatial analysis unfeasible. The report's critique lays bare the foundational data problems that must be addressed to enable more effective research and policy development in the Portuguese real estate sector:

> Data on the real estate market is not as transparent and accessible as would be desirable, which can be explained by the existence of multiple agents who benefit from this asymmetry of information. (...) it is important to raise an alert concerning the need for regulatory intervention in this area. Additionally, there is a lack of culture, rooted in public bodies, regarding the collection and supply of information, as well as the implementation of qualified, effective, and consistent decision-making processes. Although some efforts are recognized, there is still insufficient data to analyze shortages and access to adequate housing with the desirable granularity and precision, not to mention the devaluation of

prospective analyses that would allow a timely response to these needs.

– Rodrigues (2022)

A comparison of the data types presented in Table 2.4 with the range cataloged in Table 2.2 reveals a notable deficit in big data within the real estate sector. This shortfall, however, extends beyond the domain of real estate, signifying a national challenge. The INE demonstrates a delay in embracing big data, failing to integrate it into its operations and platform. This lag not only underscores a mismatch between the institute's practices and the evolving demands of contemporary data analytics but also underscores the institute's difficulty in navigating the rapidly changing data landscape (Cardoso, 2022). Additionally, a recent analysis by European Comission (2023) acknowledges some advancements in areas such as AI and blockchain in Portugal; yet, it also identifies a void in big data integration. This oversight is important because, as discussed, big data can enhance the precision, responsiveness, and effectiveness of policies in the real estate sector.

Even if big data is not yet adopted, microdata is an alternative that also allows for more complex analysis and insights.[6] The INE does provide researchers with such data. However, the real estate data provided is extremely limited and not up-to-date. Its official documentation states that it has four databases on real estate, when in fact only three are currently active, as one has been discontinued. Specifically, the active databases include: 'Construction of Buildings', with data spanning from 1994 to 2010, but with provisional data from 2011 to 2018, and only estimates from 2019 to 2021; 'Municipal Property Tax (IMI)' and the 'Municipal Tax on Onerous Property Transfers (IMT)' sourced annually from the AT. The fourth database, concerning the monthly rental value per square meter of usable area for primary residences, has been phased out. This information would be helpful to monitor the evolution of the rental market, at least on an annual basis.

[6] Microdata is detailed individual-level information that can be a component of big data, but it is not necessarily synonymous with the larger, complex datasets that big data refers to.

2.4.3 Propositions

In addressing the challenges outlined in the Portuguese real estate data landscape, the integration of big data, the use of AI and blockchain, and conducting data linkage by public institutions, present possible avenues to enhance data quality, availability, diversity, and accessibility. These solutions can help address the identified issues of data scarcity, fragmentation, and opacity.

Big Data. It can be instrumental in overcoming the limitations of data scarcity and fragmentation. By aggregating and analyzing vast amounts of data from diverse sources, big data types can provide a more comprehensive and granular view of the real estate market. A similar project to the *Platform of Indicators, Statistics, and Open Data of Justice*, identified by European Comission (2023), could be implemented for the real estate sector. The project seeks to create a comprehensive platform that collects and processes justice-related big data, merging data from judicial services and other entities. In deciding what data to collect, institutions can guide their efforts using the questions presented in Subsection 2.3.1. Even if public institutions cannot directly collect big data, they can still partner with private entities. In the real estate domain, this would mean establishing partnerships with companies like Casafari, Alfredo, and Reatia, which are already doing it, to make this data available for research and policymaking efforts. Alternatively, a concrete first step in increasing data availability and granularity could involve the AT make available (at least to the INE) its microdata on real estate and leases, as proposed by Rodrigues (2022). Similarly, since registration with this institution is mandatory, the IMPIC, which holds comprehensive data on all transactions conducted by real estate agencies, could make this information available. Implementing either of these alternatives would be a first step towards big data adoption, encouraging data innovation and sharing among real estate data providers.

AI. By automating data collection and processing of real estate data and documentation, AI can facilitate the implementation of better data management practices. It seems some strides have been made in this respect. The National Strategy for Artificial Intelligence: AI Portugal 2030, aims to advance the country's AI capabilities and align with the European Coordinated Plan on AI, with one of the goals being to improve policymaking through evidence-based decisions. Similarly, the Artificial Intelligence and Data Science for the Public Administration (AI4PA) initiative aims to facilitate the digital transition of public administration through AI to enhance the effectiveness of public policies and the quality of services. The Portuguese Recovery and Resilience Plan (RRP) funds the initiative. In the real estate sector, one of the main bodies in which such adoption would benefit the real estate sector would be the IRN. While it has adopted the technology for nationality attribution, no such effort has been made for real estate property transactions for which it is responsible.[7] This would enable processing an extensive collection of documents and data related to real estate transactions and ownership, facilitating their conversion into digital format.

Blockchain. Portugal is one of the signatories of the European Blockchain Partnership (EBP). This collaboration empowers it to harness the European Blockchain Services Infrastructure (EBSI) to advance blockchain applications within public administration and society. Additionally, Portugal capitalized on resources from the RRP to form a national consortium for blockchain comprising research centers, companies, and universities. This consortium, known as BlockchainPT, is dedicated to fostering the development of cases, pilots, and products using the technology, serving both corporate entities and public institutes. Among the consortium's participants is Unlockit, which seeks to employ blockchain technology to revolutionize real estate transactions and land ownership management. This ambition aligns with efforts in other nations, such as Ubitquity in Brazil, ChromaWay in Sweden, and ConsenSys Codefi in the UK. However, as noted, realizing blockchain's full potential in real estate necessitates a collective approach, particularly involving public registries

[7] Anecdotal evidence of this initiative can be found at https://irn.justica.gov.pt/Noticias-do-IRN/IRN-lanca-nova-plataforma-de-tramitacao-dos-processos-de-Nacionalidade.

and other public institutions. Thus, Unlockit's initiative requires support beyond funding, depending on the crucial partnership with IRN and support from institutions like the AT, ADENE, and relevant municipality halls to foster progress within the sector. This collaborative effort promises to enhance the efficiency of real estate transactions and ownership transfers and improve the collection, processing, and accessibility of such vital real estate data. The digitization of records with the help of AI, combined with blockchain, a platform for real estate transactions and ownership, would establish the foundations for a comprehensive national real estate transactions database. Such a database would likely become an invaluable resource for research and policymaking, enhancing transparency and efficiency in the sector, mimicking the examples of HM Land Registry Open Data and HDB Resale Transactions, in the UK and Singapore, respectively. Furthermore, it would be an alternative to the data monopolist Confidencial Imobiliario that uses secondary data from 'Multiple Listing Services' rather than primary data from 'Public Registries' like we are proposing. Lastly, although the potential of blockchain for facilitating open data access has been acknowledged in Subsection 2.3.3, Portugal has yet to launch an initiative specifically aimed at exploiting blockchain technology for this purpose. Utilizing the collective efforts of BlockchainPT and RRP funds to kickstart a dialogue on the feasibility of blockchain for open data in Portugal would represent a critical step toward overcoming some of the data challenges identified in the real estate sector and generally in open data.

Data Linkage. It can improve the ability to derive insights from current datasets. In particular, it can help address the identified fragmented nature of real estate datatsets and reconcile the different geographic levels. As mentioned, data interoperability between institutions is required to strive for data linkage. The Interoperability Platform of the Public Administration (iAP) stands out as a pivotal mechanism for enhancing interoperability among public institutions. It specifically enables integrating information systems across participating entities, addressing technical and semantic interoperability aspects. Notably, this platform's utility has been gradually extended to include the private sector. This should enable the linking and geocoding of datasets. In real estate, the first step in this direction would be to connect at least some public institutions involved in the sector, such as the IRN, IMPIC, AT, and ADENE. This could be done by leveraging the existing platform (iAP). Alternatively, a new one could be implemented, based on blockchain technology, to adopt the principles of 'transparency and privacy by design' and be facilitated by companies like Unlockit or other BlockchainPT members. As mentioned, the real estate industry has witnessed private institutions' creation of proprietary data collections. Casafari, Alfredo, and Reatia are such examples in Portugal. Linking data between public institutions and making such data openly available could provide alternatives to these providers.

2.5 Discussion

The enhancement of real estate data holds the potential to benefit a wide array of stakeholders within the real estate ecosystem. By expanding access to comprehensive and accurate data, individuals and organizations can make more informed decisions. This not only enhances the ability of policymakers to devise evidence-based policies that address current needs and future real estate challenges but also benefits those directly involved in the sector, namely those buying, selling, renting, investing, and developing.

The quest to enhance real estate data necessitates a unified approach, calling for collaboration across government bodies, the private sector, and academia. By adopting big data, leveraging technologies like AI, and blockchain, and promoting data sharing and interoperability between institutions, we stand to cultivate a real estate sector characterized by transparency, efficiency, and responsiveness. As evidenced in Portugal, even if mechanisms are available to back some of these ventures, the journey toward improvement extends beyond financial support. It requires a collaborative and innovative mindset across the sector.

The strategies proposed herein do not emerge in isolation. In Portugal, they are integral to ongoing national programs and legislative actions. Furthermore, they align with international initiatives, particularly at the EU level, and reflect broader trends observed in North America and other regions. The pioneering efforts seen in countries like Australia, Canada, the UK, the US, Singapore, and Sweden, and endeavors like the international #Housing2030 initiative, serve as a source of inspiration, especially in the context of a global housing crisis. These examples spotlight the importance of open data and empirical evidence in real estate, advocating for a global movement to enhance data for real estate policy. Through continued collaboration, innovation, and a commitment to open and accessible data, we can envision a future where real estate policies are informed by robust evidence, leading to more inclusive, sustainable, and responsive urban environments.

2.6 Conclusion

In this paper, we outline and introduce leading solutions from the literature for public institutions to enhance their real estate data. This is critical as addressing the challenges in real estate policymaking, particularly the housing affordability crisis, necessitates bridging the 'policy-outcome gap', which in turn depends on high-quality data. We identified four complementary approaches for public institutions in real estate data generation and distribution to enhance their data: adopting big data, integrating AI and blockchain technologies, and improving data linkage. We then focus on applying these approaches within the context of Portugal. Despite the real estate sector's importance in the national economy and the acute housing affordability crisis, Portugal appears to be lagging in employing these solutions. This is important because real estate data lays the foundations for fostering research and supporting evidence-based policy, ultimately leading to better addressing current and future real estate challenges.

Chapter 3

Machine Learning Counterfactual Imputation with Endogenous Covariates: The Pandemic's Effect on Peer-to-Peer Accommodation

Bernardo Forbes Costa, Carlos Daniel Santos, Qiwei Han

3.1 Introduction

The COVID-19 pandemic significantly setback the global economy, with the tourism sector being one of the most adversely affected. Starting in January 2020, as countries worldwide enforced lockdowns and travel bans, the ramifications for tourism became increasingly pronounced. Portugal, a nation intrinsically linked to tourism, felt this impact profoundly. In 2020, the tourism sector's Gross Value Added (GVA) shrank dramatically to 7.7 billion EUR, representing nearly half of its 2019 figure. More critically, its contribution to Portugal's Gross Domestic Product (GDP) plummeted from 8.1% in 2019 to 4.4%. This indicates that the sector's downturn was not only significant in numerical terms but also in its relative significance to the national economy (see Appendix – Figure 3.8). While there was a decline in domestic demand, the primary contraction arose from the significant drop in arrivals of foreign tourists, who account for the bulk of overnight stays in the country. Due to the numerous travel restrictions, the Lisbon Municipality saw a sharp 78% decline in overnight stays by foreign tourists from 2019 to 2020 (see Appendix – Figure 3.9).[1]

The peer-to-peer accommodation (P2PA) market was no exception. According to Statistics Portugal, the municipality suffered a 74% drop in 2020 (see Appendix – Figure 3.10). In this study, we use granular data – at unit and daily levels – for the Lisbon Municipality (henceforth "Lisbon" or "Lisbon market"). The COVID-19 pandemic offers a distinctive lens through which to explore the relationship between transient economic recessions and the dynamics of the P2PA market. Investigating this relationship is especially relevant for cities such as Lisbon, where this type of accommodation is notably prevalent. Moreover, the pandemic represents a negative demand shock. This presents a departure from much of the current literature on P2PA, which predominantly focuses on shifts in supply – either augmentative, as seen with the emergence and expansion of P2PA platforms, or restrictive, like zoning

[1]Note that these overnight stays include all accommodation types, including hotels, and not only P2PA.

policies that limit accommodation-sharing availability.

The difficulty with estimating the causal impact of such large event is the absence of a counterfactual. Neyman and Rubin (Rubin, 1974; Splawa-Neyman, 1990) attempted to circumvent this unobservability of the alternative scenario by using randomization, and other methods for observational data, like propensity score matching. A decade later, the more elaborate method of generating synthetic control groups (Abadie and Gardeazabal, 2003; Abadie et al., 2010) emerged as a way to improve further the ability to derive a counterfactual – and therefore estimate causal effects – from observational data, whenever randomization was not feasible. More recently, with the advent of "big data", and the growing availability of both quantity and quality of data, machine learning methods have been increasingly adopted due to their advantages compared to traditional linear models (Varian, 2014). Although machine learning models can lack interpretability, this is not an issue when the main goal is prediction. It turns out that recovering a counterfactual scenario is essentially a predictive task, where machine learning algorithms work best. We will refer to this approach as *machine learning counterfactual imputation*.

Machine learning counterfactual imputation offers an important advantage for our context over traditional econometric methods: it allows researchers to estimate counterfactual outcomes even in the absence of comparable historical data. This is particularly useful when all individuals are treated or affected by a policy or shock, such as in the case of a global pandemic. Unlike synthetic control groups, matching or differences-in-differences, which rely on the existence of untreated comparable groups, counterfactual imputation generates new data.[2] By modeling the outcome variable in the absence of the policy or shock and comparing it to the observed values, researchers can quantify the policy or shock's impact. This approach allows for time-varying confounders that may affect the outcome of an intervention. These

[2]An alternative naive approach would be to consider a before-after estimator. While similar in spirit, this method is susceptible to biases from time-varying unobserved confounders and fails to capture nonlinear relationships between variables, potentially yielding misleading conclusions about the event's true impact.

variables can be modeled as predictors of the outcome, improving the accuracy of the estimated counterfactual model. This approach is relatively straightforward to implement, requiring only a predictive model for the outcome variable.

The machine learning counterfactual imputation approach popularized by Varian (2016) produced a growing body of literature that estimates treatment effects in the absence of a control group. More specifically, it has been used to study the effects of COVID-19 on the energy sector (Benatia and Gingras, 2023; Fabra et al., 2022), on general macroeconomic activity (Cerqua and Letta, 2022; Dueñas et al., 2021), and mortality estimates (Cerqua et al., 2021). We use this methodology to study the impact of the COVID-19 pandemic (our "shock" or "event") on the P2PA market for Lisbon, more specifically on Airbnb, the world's largest home-sharing platform and the prevalent one in Lisbon.[3] It should be noted that the term "shock" refers to the economic consequences of the pandemic, encompassing not only the direct spread of the virus but also the associated changes in behavior and the implementation of local and international interventions, including travel bands and lockdowns. We define the pre-treatment or "pre-pandemic" period as the time horizon before March 2020 and the post-treatment or "pandemic" period as any subsequent period (i.e., after and including the 1st of March 2020), which in our dataset corresponds to the rest of the 2020 year. Even if the "State of Emergency" in Portugal was only declared on the 18th of March, the first case of COVID-19 was confirmed on the 2nd of March 2020.

As we cannot obtain an adequate control group, as all listings were treated, we apply machine learning counterfactual imputation to uncover the causal effects of the pandemic. Our general approach is as follows: First, select the best-performing machine learning model and train it on the pre-treatment data. Second, deploy it to

[3] As in many European countries, Airbnb is not the only short-term rental platform active in Portugal. However, it is the largest player in the market. According to AirDNA more than 90% of the P2PA properties in the Lisbon Municipality are listed on Airbnb. Furthermore, most P2PAs are advertised through more than one platform. This makes it highly likely that most properties will be listed on the Airbnb platform. Hence, we consider Airbnb listings a good proxy for the P2PA market in Lisbon.

the post-treatment period. Third, in the post-treatment period subtract observed outcome values from predicted ones (the imputed counterfactual) and obtain a distribution of treatment effects, which can then be averaged to derive the average treatment effect on the treated (ATT).

This methodology is related to the work of Athey et al. (2021) who use a matrix completion algorithm to generate the counterfactuals, Borusyak et al. (2023) who define and use linear fixed effects "counterfactual imputation", and Gobillon and Magnac (2016), and Liu et al. (2022) who use interactive fixed effects. Closest to our approach is Souza (2022) who uses a machine learning counterfactual imputation and shows through simulations that machine learning algorithms can enhance the accuracy of counterfactual predictions, as evidenced by reduced out-of-sample errors, optimizing the estimation of treatment effects.[4]

However, the literature has ignored one important element: some of the covariates may be endogenous. For example, Airbnb's prices are an important predictor of bookings but they were (endogenously) lowered during the COVID-19 pandemic (Milone et al., 2023). All of the proposed methods work under the assumption of pre-determined or exogenous covariates.[5] That is, they assume that covariates are unaffected by the treatment. For instance, Borusyak et al. (2023) requires covariates to be unaffected by treatment and strictly exogenous to be included in their counterfactual imputation specification. We relax this fundamental assumption. Fabra et al. (2022) suggest that if demand is price-elastic, the counterfactual demand predictions would need to factor in potential price shifts stemming from the analyzed policy or event, but without specifying how this should be achieved. Souza (2022) recognizes the problem but states it is out of the scope of his study. Nevertheless, he hints that a possible solution would be to predict counterfactual realizations of

[4]In contrast to *in-sample* analysis, which evaluates a model on the training data, *out-of-sample* refers to the evaluation of a separate, unseen portion of the data, providing a measure of the model's generalizability and predictive capabilities beyond the training data.

[5]This problem has also been discussed in the matching literature (Heckman and Navarro-Lozano, 2004; Wooldridge, 2016).

these covariates. We follow his trail of reasoning.

Our example is straightforward. We want to predict P2PA host revenues in the absence of a proper control group to estimate the impact of the COVID-19 pandemic. However, in our context, one of the most important predictors of revenues (price) is endogenous. That is if we would use as a covariate the prices that were observed under the treatment – which have adjusted and were, for example, lower than the counterfactual prices – this would deliver severely biased estimates for the effect of the pandemic on the revenue, even if the treatment (pandemic shock) is exogenous. Instead, we propose a method to perform counterfactual imputation in the presence of endogenous variables. Our approach requires the use of multiple (in our case two) machine learning models: one auxiliary model for the endogenous covariate and one main model for the variable of interest (revenue). This is intuitive and bears a resemblance to the standard instrumental variable's (IV) literature, although it does not require any variable satisfying a standard exclusion restriction. In Section 3.3, we formalize our estimation approach.

Our study contributes to the existing literature on P2PA by offering a perspective on how this market reacts to a negative demand shock in a specific geography (Lisbon). Furthermore, we add to the literature on machine learning counterfactual imputation by expanding the methodology to consider endogenous covariates to estimate causal effects.

The remainder of our chapter is structured as follows: in Section 3.2, we survey the literature on the impact of COVID-19 on P2PA globally and then offer an overview of the P2PA market in Lisbon. In Section 3.3, we formalize our empirical approach, namely our extension to consider endogenous covariates within the machine learning counterfactual imputation framework. In Section 3.4, we present our results, in Section 3.5 we discuss them, and in Section 3.6 we conclude.

3.2 P2PA and the Pandemic

3.2.1 The Pandemic Impact on the P2PA Sector

The global outbreak of COVID-19 significantly disrupted the tourism sector, particularly affecting the hospitality industry. Duro et al. (2021) identifies some of the key factors contributing to the sector's susceptibility including the country's reliance on tourism, the structure of the market, the availability of accommodations in rural areas, and the overall impact of the health crisis. Within the hospitality industry, P2PA and traditional hotels experienced different impacts. Gyódi (2022) comparative analysis of Airbnb and hotels across nine European cities offers insightful contrasts in their resilience and strategic responses during the pandemic. The study reveals Airbnb's more flexible supply. While hotels approached their 2019 occupancy rates relatively swiftly, Airbnb listings lagged, maintaining an average count 15% lower than the previous year. Conversely, Airbnb's prices demonstrated a more tempered decline compared to the steeper rate reductions observed for hotels. Gerwe (2021) offers further insights into the strengths and weaknesses, as well as recovery prospects of P2PA. The author discusses the potential for a "reset" in the post-pandemic recovery phase, which implies a rethinking and reorientation of strategies within the accommodation-sharing sector.

The pandemic not only highlighted differences between traditional hotels and P2PA but also significantly influenced consumer preferences. This period saw a pronounced shift in demand towards more rural, less dense, and individual lodging options, marking a distinct change in the landscape of travel and accommodation preferences. Zoğal et al. (2020) explored the role of second homes during and after the pandemic, noting a growing preference for short-term rentals in less densely populated areas. This shift is largely attributed to the challenges of maintaining social distancing in urban environments, which led travelers to seek accommodations in more secluded locations. Similarly, Jang et al. (2021) analyzed the impact of the pandemic on different destination types in Florida, highlighting significant variations in Airbnb revenue losses between urban and rural areas. The study also observes that business tourists, less concerned about the pandemic, demonstrated a higher propensity to use Airbnb compared to leisure tourists, emphasizing the need to consider traveler types and their risk perceptions. Bresciani et al. (2021) delve into the psychological drivers behind accommodation choices during the pandemic. The study observes a marked preference for full flats over shared spaces and hotels, a trend driven by the desire for physical distancing. This preference signifies a broader change in consumer behavior in response to health concerns. Lastly, Sainaghi and Chica-olmo (2022) provides insight into the changes in Airbnb rental patterns in Milan, where the pandemic reduced the appeal of central city locations in favor of more peripheral areas. This shift indicates a clear preference among travelers for less crowded spaces, a trend reflective of the broader shifts observed in the sector during the pandemic.

Zhang et al. (2021) and Chen et al. (2021) provide complementary insights into how PSPA hosts have been impacted by COVID-19. Zhang et al. (2021) explore the responses of Chinese hosts to the pandemic, categorizing them based on their strategies. The study raises the question of whether COVID-19 has been a terminator or an accelerator for P2PA. The findings suggest that the pandemic catalyzed the P2PA sector, favoring hosts who adhered to the core principles of peer-to-peer accommodation while filtering out those motivated primarily by profit. On the other hand, Chen et al. (2021) focuses on the economic vulnerability of the sharing economy, specifically through the lens of Airbnb hosts in Sydney, Australia. The study reveals a substantial 89.5% drop in income for these hosts between January and August 2020, amounting to an approximate loss of 14 million AUD. This significant financial impact, which was 6.5 times greater for hosts than for Airbnb itself, underscores the sharing economy's susceptibility during the pandemic.

Cheng et al. (2023) provide a comprehensive analysis of the global impact of the pandemic on Airbnb, focusing on the consequences of the initial Wuhan lockdown, local COVID-19 case numbers, and local lockdown policies. Their study reveals a significant 57.8% decrease in global Airbnb bookings due to local lockdowns and a 4.16% decline in bookings with every doubling of COVID-19 cases. The authors find that the impact of the shock on bookings lessens with increasing geographic distance from Wuhan and intensifies with stricter government lockdown policies and higher levels of human mobility within a market. To derive these insights, the authors employ a DiD approach, comparing Airbnb booking activities before and after the Wuhan lockdown, with the year 2019 serving as the baseline period. Additionally, the study leverages geographical variations in the stringency of local COVID-19 responses and cases to study the sensitivity of booking activities to the different pandemic factors. Similarly, Milone et al. (2023) delve into Airbnb pricing and demand (i.e., bookings) dynamics in Europe during the pandemic. To address the endogeneity of demand and prices, the authors utilize an IV approach, which leverages variations in country-level policy responses to the pandemic. Concretely, in the first stage, the study instruments the endogenous variable, Airbnb bookings, with the COVID-19 Stringency Index. In the second stage, the estimated Airbnb bookings are then used to measure the impact of demand on Airbnb prices. The study found that a one percent increase in the COVID-19 Stringency Index is associated with a reduction of 0.133% to 0.155% in the number of bookings. Conversely, a one percent rise in the number of bookings is associated with an increase in prices ranging from 0.06% to 0.07%.

Overall, these studies show that the COVID-19 pandemic had a significant impact on the P2PA sector, leading to decreased bookings and income loss for hosts, showcasing its vulnerability to shocks. Plus, the impact of the pandemic seems to depend on local characteristics, namely the stringency of local measures adopted. Our contribution to this body of literature involves estimating the pandemic's causal effect on Airbnb host revenue in the Lisbon P2PA market. We employ a machine learning counterfactual imputation methodology, described in detail in Section 3.3. While our approach, like Milone et al. (2023), addresses the endogeneity of prices and bookings, it uniquely focuses on the specific geography of Lisbon rather than relying on country-wide variability. This allows for a more detailed estimation of the pandemic's effect within this particular context. Our study parallels Batalha et al. (2022) in terms of institutional framework, although their focus is primarily on housing prices rather than the direct impact on the P2PA market. Next, we delve into characterizing the P2PA market in Lisbon.

3.2.2 The P2PA Lisbon Market

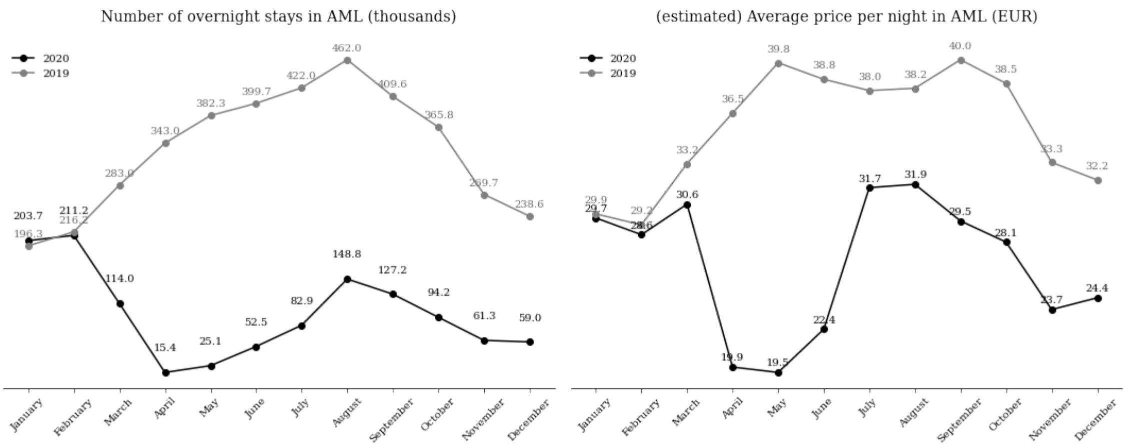
According to Statistics Portugal, in 2019, Lisbon accounted for a significant portion of Portugal's tourism sector: 20% of the nation's total overnight stays, and 31% of those in P2PAs nationwide. The accommodation business in Lisbon has experienced impressive growth. From 2011 to 2019, total overnight stays skyrocketed from 6.4 million to 13.9 million (see Appendix – Figure 3.9). This surge was predominantly fueled by international visitors, registering a 143% increment, although domestic demand too saw a respectable increase of 41%. The trajectory for P2PA in the region was similar, witnessing a 139% surge between 2017 and 2019 (see Appendix – Figure 3.10).

This is no coincidence. Firstly, Lisbon has become a very attractive destination for tourists, with the World Travel Awards recognizing it as the top city for short vacations for four straight years from 2017 to 2020. Secondly, a nationwide regulation change that facilitated the listing of properties as P2PA (*Decree-Law no. 128/2014*) resulted in a major increase in the market supply. Adding to that, a 2016 Airbnb report for the Lisbon market stated that 30% of guests would not have visited Lisbon or stayed longer had they not had access to Airbnb. Hence, Lisbon became a popular destination for tourists, P2PA supply increased drastically and this type of accommodation seems to be an important driver of tourism demand. This naturally resulted in a sudden increase in property prices (Franco and Santos, 2021), which led Lisbon's local authorities to curb new registrations of short-term rentals in certain areas of the municipality (*Proposal 677/CM/2018*).[6] This restriction was then broadened to include additional areas in 2019 (*Proposal 204/CM/2019*).

However, this growth trend in P2PA suffered a major setback in 2020. Figure 3.1, for the Metropolitan Area of Lisbon (AML), depicts the distinct contrast between the pre-pandemic and pandemic periods with more detail. Given that most of the P2PAs for the AML are in Lisbon (79% in 2019), we expect the identified patterns to hold for our geographical area of analysis. One prominent trend is the uptick during the summer months of July through September, evident in both the volume of overnight stays and the average pricing. This trend persists in both 2019 and 2020. Another notable pattern is the congruence in values for January and February across both periods, only to witness a sharp divergence in March 2020 – a testament to the pandemic's onset. This shift resulted in diminished demand for accommodation, consequently driving down prices. Notably, in July and August, the disparity between the pre-pandemic and pandemic figures for both pricing and overnight stays narrowed substantially. This reduction in disparity is indicative of the relaxation of travel restrictions aligned with the summer season, ushering in a renewed surge in demand. However, this narrowing trend was short-lived, as rising COVID-19 cases and the resultant ramping up of containment measures began to exert their influence as the year progressed (see Appendix – 3.9.1). The result, in the opposite direction of the positive supply shock that the 2014 policy generated, was a decrease in housing prices (Batalha et al., 2022).

[6] These areas were denominated *Zonas Turísticas Homogéneas* and had more than 25% of the dwellings registered as P2PA.

Figure 3.1: Number of overnight stays and estimated average price for P2PAs



Note: Data from Statistics Portugal. The average price per night is computed by taking the ratio of the revenue and the number of overnight stays for each month. This figure analyzes data at the geographical level of the Metropolitan Area of Lisbon (AML), offering a broader scope than the Lisbon area alone.

3.3 Empirical strategy

In this section, we describe our approach to using machine learning methods to recover a counterfactual scenario of a world without the effects of the COVID-19 pandemic. Our goal is to compare estimates of the unobserved outcomes with actual observations to infer causal effects and evaluate the impact of the pandemic on P2P in Lisbon. In doing that we address the issue of having as endogenous an important predictor for our machine learning counterfactual imputation model - prices. Dealing with the endogenous response of prices is fundamental given the observed price changes. Our counterfactual scenario using actual (lower) prices during the pandemic would underestimate revenues when compared with the counterfactual scenario for prices under no pandemic.

3.3.1 Machine Learning Counterfactual Imputation

We borrow our conceptual framework from Fabra et al. (2022) who study the impact of COVID-19 on carbon emissions using the same machine learning counterfactual imputation methodology we apply here. Building on the potential outcomes framework by Neyman-Rubin, the goal is to obtain the average treatment effect. In our case, the treatment is the COVID-19 shock, for which there is no reliable comparison group because all units were treated. Let i represent a single Airbnb listing, $t = pre$ the set of days $t \in [-403, -1]$ representing the pre-pandemic period, and $t = pand$ the set of days $t \in [0, 305]$ indicating the pandemic period, where $t = -403$ and $t = 305$ are the first and last dates in our dataset (January 23$^{\text{rd}}$, 2019, and December 31$^{\text{st}}$, 2020), respectively, and $t = 0$ the start of our pandemic period (March 1$^{\text{st}}$, 2020).

Further, let $Y_{it}(p)$ represent either (i) the price or (ii) revenue (both in EUR), for listing i at time t, and $X_{it}(p)$ the respective set of covariates, both depending on the binary variable p, which denotes whether the realization was affected ($p = 1$) or not ($p = 0$) by the pandemic shock.

The challenge is to obtain the counterfactual scenario or $Y_{i\,pand}(0)$ to which we can compare our actual treatment results. We do this by building a model to estimate $Y_{i\,pand}(0)$, using our pre-pandemic sample which leverages the following relationship:

$$Y_{i\,pre}(0) = M(X_{i\,pre}(0)) + \varepsilon_{i\,pre} \tag{3.1}$$

$$\text{such that } \mathrm{E}[Y_{i\,pre}(0)|X_{i\,pre}(0)] = M(X_{i\,pre}(0))$$

where $M()$ is a generic function or model (e.g., linear regression, random forest, boosting or neural networks). We add to this structure the traditional assumption in event studies that agents did not change their behavior in anticipation of the shock:

Assumption 1 *No anticipatory effects*

$$Y_{\text{pre}} = Y_{\text{pre}}(0).$$

This is a credible Assumption given that the COVID-19 pandemic was completely unexpected. Furthermore, we require that the relationship established in equation (3.1) holds for all t, such that it is valid for both the *pre* and *pand* periods. This allows us to rewrite it as follows:

Assumption 2 *Stability of the counterfactual function*

$$Y_{i\,pand}(0) = M(X_{pand}(0)) + \varepsilon_{i\,pand} \quad (2) \tag{3.2}$$

such that

$$\mathbb{E}[Y_{\text{pand}}(0)|X_{\text{pand}}(0)] = M(X_{\text{pand}}(0)).$$

This Assumption is also credible in our setting because there were no other changes to the short term rental market during this period. Finally, to identify the counterfactual we require the selected covariates to be independent from the COVID-19 shock. Nevertheless, we have the benefit of not requiring them to be independent from the error term because that happens naturally. In other words, treatment is unrelated with the listings' characteristics:

Assumption 3 *Covariates are independent from treatment (COVID-19 pandemic)*

$$X_t(0) = X_t(1) = X_t. \tag{3.3}$$

Note that this assumption does not preclude us from using realizations of covariates X_t during the pandemic period. On the contrary, the ability to use future values of predictors is one of the main advantages of this approach. By combining Assumptions 1, 2, and 3 we can rewrite equation (3.2) as follows:

$$E[Y_{pand}(0)|X_{pand}] = M(X_{pand}). \tag{3.4}$$

Thus, we can identify the machine counterfactual and conduct the following estimation procedure:

- Estimate our model based on the pre-pandemic data:

$$Y_{i\,pre} = M(X_{pre}) + \varepsilon_{pre} \tag{3.5}$$

such that

$$\hat{Y}_{\text{pre}} = \widehat{M(X_{\text{pre}})}.$$

- Build the machine learning counterfactual based on the previously estimated model, using pandemic data:

$$\hat{Y}_{pand} = \widehat{M(X_{pand})}. \tag{3.6}$$

The main assumptions for this approach to be valid are that the counterfactual function is stable and there are no anticipation effects. In our context both are plausible. Assumption 1 should hold for the following reasons: We are using data from the pre-pandemic period, while the exact timing and consequences of the pandemic were unknown ex-ante. Furthermore, the full-on consequences of the pandemic for P2PA, in the form of travel bans and lockdowns only started to take place two weeks into our pandemic period (after 18$^{\text{th}}$ of March 2020) as the Portuguese government started to respond to the increasing number of cases locally and globally. To evaluate Assumption 2, we use time-series cross-validation (Hyndman and Athanasopoulos, 2021) which should attest to our model's stability over time. This procedure is described in Appendix – 3.8.1. As for Assumption 3, we selected our covariates, from the available list of features for each listing (e.g. number of bedrooms), based on their independence with the pandemic shock (see Appendix – 3.7.2). Plus, we also use day-of-year, day-of-week, and month-fixed effects, which are exogenous by definition.

3.3.2 Endogenous Covariates

We are interested in predicting prices and revenues for each listing in a scenario in which there is no shock. We do this by obtaining individual forecasts at a daily level: a forecast for each listing for a specific day. Nevertheless, our forecasting models are pooled, which implies that while we have separate predictions for each listing, every approach employs an identical model for all listings. By pooling our models, we avoid overfitting and simultaneously capture local cross-listing dynamics. Both these models use covariates which are independent of the pandemic shock.

The issue is that one of the most important predictors of revenue, price, is endogenous. If we directly use price as a predictor, we will be violating Assumption 3 and will obtain significantly biased estimates of the treatment effect. We call this *endogeneity bias*. Concretely, if we model revenues on pre-pandemic data ($t = pre$) where one of its most important predictors is price, and then use this model to predict revenues in the pandemic period ($t = pand$) where prices have adjusted (for example lowered) in response to the shock, this could result in lower predicted revenue, thus underestimating the effect of the pandemic.[7] We test for this and report the results in Subsection 3.4.3. Whereas adding exogenous covariates does not create any bias, introducing endogenous ones will severely bias our estimates, particularly for high-complexity machine learning models.

To obtain the correct counterfactual and avoid violating Assumption 3 we can take two alternatives. First, we drop the offending variable (price) as a predictor of revenue and estimate the counterfactual abstracting from this offending variable. However, given the importance of price as a predictor, this will lead to a deterioration of the predictive ability of our model (see Appendix – 3.8.7). The second, and the one we adopt, is similar to the first stage of an instrumental variable approach. We first use a model to predict the endogenous covariate and replace the endogenous

[7] Note that by directly predicting revenue we are also implicitly modeling booking status given that if a booking has zero revenue it is because it was not booked.

covariate with its predicted value. Since we are not concerned with identifying the effect of the endogenous covariate, we do not require any exclusion restrictions. This first stage's objective is to improve the accuracy of the predicted counterfactual from the first step.

There are three potential concerns with this approach. First, the "instrumentation" of the endogenous price variable has no instruments that satisfy the exclusion restriction in the first step – it must not be directly correlated with the error term in the main regression equation. All covariates are also present in the main outcome equation. This is not a problem as long as we are not interested in identifying the effect of the endogenous covariate. For example, in our case, we are not interested in obtaining an estimate of the price elasticity. We also have to be careful not to include an instrumental variable and the endogenous regressor in the main equation of interest (Wooldridge, 2016). To avoid this we include the price predicted by the counterfactual model instead of the observed value. Second, machine learning algorithms, like the random forest regressor that we use, may introduce some bias to reduce variance, typically called *regularization bias*. To address this Chernozhukov et al. (2018) propose a debiased estimator using sample splitting (cross-splitting). We adopt the sample splitting (cross-splitting) procedure – which in our case is done with time-series cross-validation (see Appendix – 3.8.1). Third, by using the predicted price from the first stage as a predictor in our revenue model, we may still have a bias if there is some form of endogeneity still left in the predicted prices due to a violation of Assumption 3. To address this, we adapt to our case a similar control function approach proposed in Heckman and Navarro-Lozano (2004) to solve the residual endogeneity problem due to omitted conditioning variables.[8] We thus use the predicted value of the endogenous variables (instead of the actual value) together with the residual from the first stage as covariates in our main model. Specifically, we collect the out-of-sample residuals–from the time-series cross-validation procedure–

[8]The control function solution to solve the endogeneity problem is to use as covariates in the main equation the endogenous variable and the predicted residual from the first stage.

for the price model, compute their average value for each listing, and add that as a predictor to our revenue model. On the one hand, time-series cross-validation residuals offer a more robust and generalizable basis for capturing the true price residuals, free from overfitting biases. On the other hand, by using the average of these out-of-sample residuals, we align our approach with the principle by Blundell et al. (1999), where constructing a fixed effect from a pre-sample helps to control for unit-specific trajectories and unobserved heterogeneity. This is especially pertinent in the context of counterfactual imputation, where the objective is to forecast potential outcomes based on a model trained with historical data. The average residual, in this sense, serves as a proxy for unobserved factors that consistently influence the revenue potential of a listing across time, beyond the observable characteristics and the directly modeled effects of price changes.

In Appendix – 3.8.7 we show that using the residuals as a predictor, not only reduces the bias of the main model – as given by the difference between the mean out-of-sample residuals for a revenue model with and without the predictor – but also improves its predictive performance. We also show that, as a result of our approach, bias is negligible for both the price and revenue models as given by the out-of-sample residuals (see Appendix – 3.8.8). This is important because Souza (2022) shows that out-of-sample (cross-validated) residuals approximate well the true counterfactual residuals – which are by definition unobserved. Finally, our chosen model performs significantly better than the linear regression (see Appendix – 3.8.2), which should not be affected by regularization bias.

Given the classical demand and supply dynamics in which the price (charged by hosts) affects the demand (revenue generated if booked), we train the models sequentially and adopt the following five-step approach to build the machine learning counterfactual:

1. **Pricing model**: we first build our model for prices according to equation (3.5)

and obtain an estimate for:

$$\hat{P}_{i\,\mathrm{pre}} = \widehat{M_1(X_{i\,\mathrm{pre}})},$$

2. **Residual estimation**: we apply time-series cross-validation and obtain the average out-of-sample prediction residual for each listing:

$$\hat{\epsilon}_{i\,\mathrm{cv}} = \frac{\sum_{j=1}^{J}(\hat{P}_{i\,\mathrm{pre}} - P_{i\,\mathrm{pre}})}{J},$$

   where J is the number of times (number of dates) that each listing shows up in the testing set (i.e., out-of-sample). Note that we obtain one single value (i.e., a constant) for $\hat{\epsilon}$ for each listing for the pre-pandemic period.[9]

3. **Revenue model**: we consider the same covariates used in the pricing model but extend them with the estimates for pricing and residual obtained in the previous steps:

$$\hat{R}_{i\,\mathrm{pre}} = M_2(X_{i\,\mathrm{pre}}, \widehat{\hat{P}_{i\,\mathrm{pre}}}, \hat{\epsilon}_{i\,\mathrm{cv}}).$$

   Notice the coefficients on covariates X for M_1 and M_2, in our case, are "nuisance" parameters. The advantage of following this two-stage procedure is that it allows for the introduction of a predictor that otherwise would be endogenous – given that prices were affected by the shock, but the predicted ones are not. This is particularly important given that price is an important predictor of revenue.

4. **Machine learning counterfactual imputation**: after training both the pricing and revenue models, sequentially and in this order, with the entire pre-pandemic data, we predict, for the pandemic period, prices and revenues,

[9]Given the structure of our time-series cross-validation approach some listings might not show up in the testing set. For these, we assume a predictive residual of zero.

using the models trained in steps **1** and **3**, respectively. These estimates serve as our machine learning counterfactual for prices and revenues.[10]

5. **Causal effect**: compute the difference between observed and counterfactual revenue values for the pandemic period and obtain an estimate of the effect:

$$\bar{\alpha}_{\text{pand}} = \frac{\sum_{i=1}^{N_{\text{pand}}} \sum_{t=0}^{T_{\text{pand}}} (R_{i\,\text{pand}} - \hat{R}_{i\,\text{pand}})}{T_{\text{pand}} \times N_{\text{pand}}}, \tag{3.7}$$

where T_{pand} is the number of pandemic days, N_{pand} the number of listings considered in the pandemic period, $R_{i\,\text{pand}}$ is the daily revenue observed during the pandemic, $\hat{R}_{i\,\text{pand}}$ the daily counterfactual revenue, and consequently $\bar{\alpha}_{pand}$ our average causal effect or ATT. Naturally, this effect can be adapted to be divided by the average $\hat{R}_{\text{pand}}$ and presented as a percentage.

Based on the comparison of cross-validated mean-squared error and R-squared, the selected algorithm was the random forest regressor for both prices and revenue models. The full model selection, hyperparameter tuning, and training procedure are described in Section 3.8 of the Appendix.

3.4 Results

We follow the five-step approach described above to impute a machine counterfactual for the pandemic period and analyze what would have happened in the absence of the shock to the listings that were active in the pre-pandemic period. This means that while we account for listings leaving the platform for the observed values, we also consider the days during the pandemic period in which a listing would have been active, but was not due to the pandemic for the counterfactual. For instance, if a listing was active on the 1st of May 2019, but was inactive on the 1st of May

[10]For model deployment to the pandemic period, we use the average out-of-sample residuals from training (i.e., from the pre-pandemic period), defined in step **2**.

2020, we still generate a counterfactual for this listing as if it was active on this date – while assuming that the property was not booked and therefore generated no revenue. This way we are capturing what would have happened to each listing if there were no exits from the platform, without disregarding the dates on which the listings operated. In doing so, we are also able to segregate new listings (i.e., new supply) that appeared during the pandemic period and evaluate their magnitude in the market.

Having imputed the machine learning counterfactual for prices and then revenues, we can compare observed and counterfactual revenues and identify the causal effect of the pandemic on the Lisbon P2PA market. The following subsections describe the impact of COVID-19 on prices, quantities, and finally on revenues to identify the causal effect of interest $\bar{\alpha}_{pand}$.

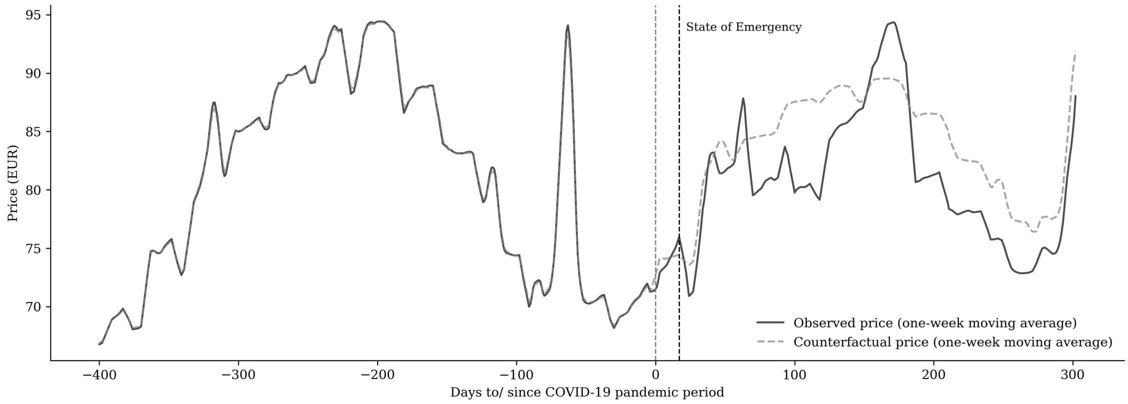
3.4.1 Prices

The average price for each listing before the pandemic was 80.72 EUR. The mean of the machine learning counterfactual imputation for the pandemic period was higher at 83.46 EUR, representing an increase of 3.4%.[11] However, we observe that the price set by the hosts was lower than expected during the pandemic, with an observed average of 80.63 EUR. This represents a slight decrease from the pre-pandemic period and a -3.4% difference from the counterfactual value. If we consider the median prices instead, which by definition are less sensitive to extreme values, the reduction in prices becomes more evident: 68.00 EUR and 65.00 EUR in the pre- and post-pandemic periods, respectively. This indicates that Airbnb hosts, on average, responded to the shock by decreasing the price charged per night to face the decreasing demand.

In Figure 3.2 we plot the machine learning counterfactual (i.e., prices estimated with our "first-stage" auxiliary model) and the observed prices for both the pre-

[11] This suggests the model picked up a positive trend in prices, which might, for instance, reflect inflation.

Figure 3.2: Machine learning counterfactual imputation and observed values – Price



Note: Observed and counterfactual prices for active listings during the pre-pandemic period. Prices displayed represent the one-week moving average centered on that day. Counterfactual prices for the pre-pandemic period represent in-sample forecasts and for the pandemic period out-sample predictions. State of Emergency represents the 18th of March 2020. While Figure 3.1 showcases data specific to the AML region, in this figure, our focus shifts exclusively to the Lisbon area.

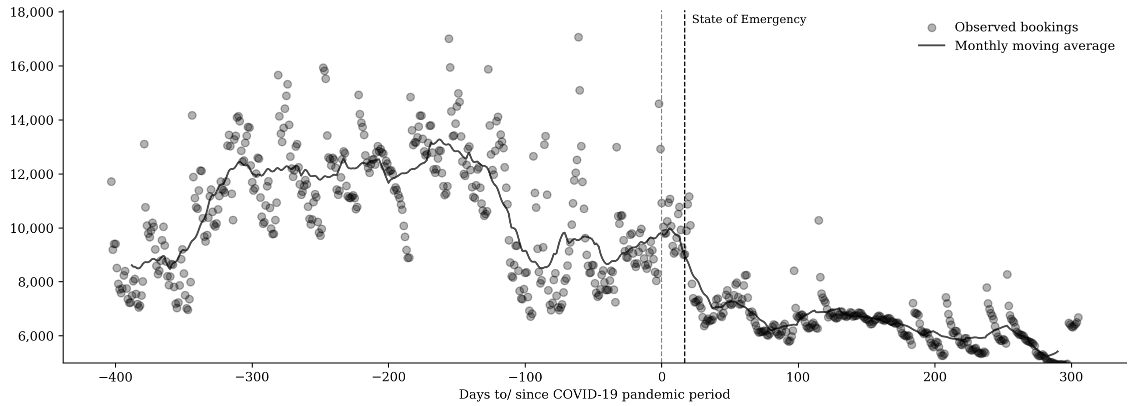
pandemic and pandemic periods. Firstly, note that for the pre-pandemic period, the observed and counterfactual prices are almost indistinguishable attesting to the good fit of our model. Secondly, the average observed prices are generally lower than the observed ones for the pandemic. Thirdly, similar to the pattern observed in Fig 3.1, there is a convergence between observed and counterfactual prices in the middle of the pandemic period reflecting the summer season and the decreases in the stringency of containment measures.[12] Fourthly, there was a sudden drop in prices following the State of Emergency suggesting hosts immediately reacted to the announced measures. Finally, there was a spike towards the end of the pandemic period which corresponds to the Christmas period when mobility drastically increased in the country, as families traveled to see their relatives and tourists took advantage of the holidays to visit the city — which naturally resulted in a surge in the number of cases in the weeks that followed.

[12]Note, however, that the geographic level of analysis here is Lisbon, rather than AML as in Figure 3.1.

3.4.2 Quantities

While we do not directly model the number of bookings, the quantity dimension is still implicit when we predict revenues directly.[13] The total number of bookings for the pre-pandemic was 4.3 million. Considering we assume a fixed supply, that is, no new listings, the figure for the pandemic period was 2.0 million. Figure 3.3 shows there was a clear downward trend that lasted until the end of the time horizon, suggesting the effect was long-lasting. Moreover, and similar to prices, there is a sudden drop following the State of Emergency declaration. This evidence suggests that demand, as given by the number of bookings for the same supply, dropped drastically.

Figure 3.3: Number of bookings



Note: Number of bookings for active listings listings during the pre-pandemic. The line displayed represents the one-month moving average centered on that day. State of Emergency represents the 18$^{\text{th}}$ of March 2020.

3.4.3 Revenues

We now turn to our main results. The total observed revenue during the pandemic amounted to 159.1 million EUR, while the counterfactual one was 280.2 million EUR. This represents a market contraction of 43.23%. The average observed daily revenue before the shock was 48.46 EUR. During the pandemic, the observed value

[13]We opted for a direct prediction of revenues rather than a separate prediction for prices times the probability of booking because of the better out-of-sample performance of the first approach (see Appendix – 3.8.3).

stood at 25.33 EUR. However, the counterfactual average daily revenue was 51.68 EUR – following the upward trend also verified in counterfactual prices.

We plot in Figure 3.4 the total (3.4a) and average (3.4b) daily values for observed and machine learning counterfactual imputation revenues. The conclusions are the same for both plots. First, our models are well-fitted to our data with the two lines for the pre-pandemic period being indistinguishable. Second, before the declaration of the State of Emergency, there is an overlap between the counterfactual and observed revenues. This indicates that the profound impacts of the shock commenced precisely on that date. This aligns with the patterns observed for prices and quantities: post-declaration, the observed revenue sharply declined, whereas the counterfactual maintained its upward trajectory. Third, the Christmas season uptick is also present here.

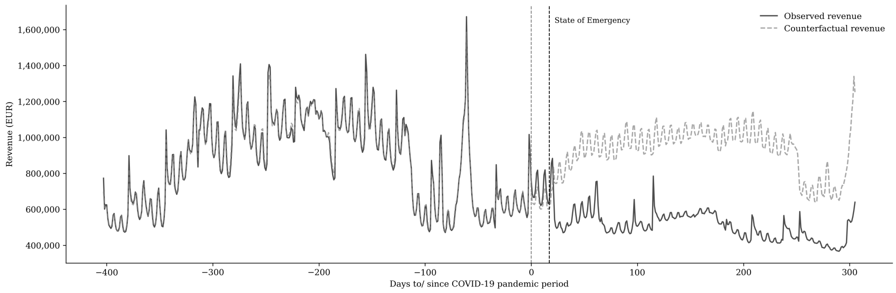
All in all, we estimate an ATT ($\bar{\alpha}_{pand}$) of the pandemic of -26.92 EUR per day for each listing. Considering that the mean counterfactual revenue for the pandemic period was 51.68 EUR, this represents an overwhelming ATT of -52.10%.[14] Assuming an average month is comprised of 30 days, hosts experienced an average monthly loss of 808 EUR. This loss is substantial, especially when compared to a reference value such as the Portuguese minimum wage of 635 EUR in 2020. Such a substantial decline emphasizes the financial challenges faced by hosts, as their losses even exceeded the standard monthly earnings of many workers in Portugal for that year. What's more, these results probably reflect a lower bound of the true effect of the pandemic. This is so because we assume that unavailable listings are booked, when in fact some might simply not be available for other reasons rather than being booked (ex. host decides to temporarily "blackout" the property). Thus, we might

[14] Attentive readers may be tempted to draw parallels between the significant effect discussed here and the 78% decrease in overnight stays in Lisbon from 2019 to 2020, as introduced in Section 3.1 and depicted in Figure 3.9. However, such comparisons are not entirely straightforward. This is because overnight stays not only encompass P2PA but also include hotels and other types of lodging. Therefore, the observed discrepancy can be attributed to the differences in how each accommodation type experienced the downturn. Specifically, Airbnb's pricing showed a more moderate reduction in comparison to the more pronounced price cuts offered by hotels, as detailed by Gyódi (2022).

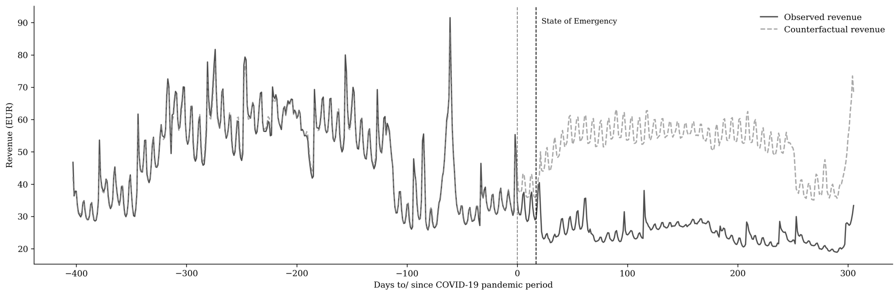
be overestimating counterfactual revenue and therefore underestimating the effect of the pandemic.

Figure 3.4: Machine learning counterfactual imputation and observed values – Revenue

(a) Total revenue



(b) Average revenue

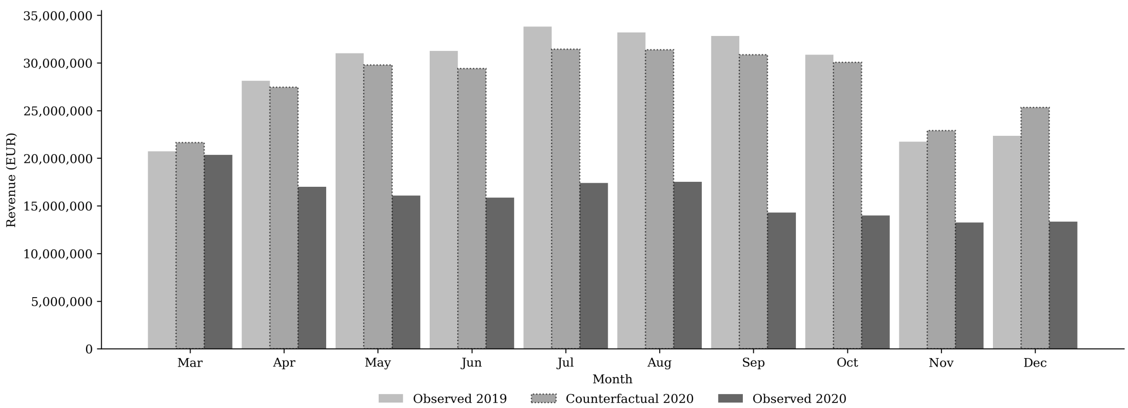


Note: Observed and counterfactual revenues for active listings during the pre-pandemic. Counterfactual revenues for the pre-pandemic period represent in-sample forecasts and for the pandemic period out-sample predictions. State of Emergency represents the 18^{th} of March 2020.

Two important notes. First, the average revenue daily loss is positively correlated with the increase in the number of COVID-19 cases and negatively correlated with the stringency of the containment measures imposed by the Portuguese government (see Appendix – 3.9.1). These correlations further substantiate the credibility of our findings as they are also found in the literature (Cheng et al., 2023; Milone et al., 2023). Second, if we follow the naive approach to predicting revenues using observed prices as predictors – as opposed to our five-step approach discussed in Subsection 3.3.2 – we would obtain an ATT of -17.07 EUR or -33.58%. This substantially underestimates the effects we obtain because it considers prices that have adjusted to the shock (i.e., decreased). This highlights the need to adopt our approach to control for endogenous covariates.

An alternative method of looking at our results is by comparing the monthly revenues during the pandemic months to their corresponding values in 2019, alongside the counterfactual for 2020. This is what is depicted in Figure 3.5. While the revenue difference between observed values in March is negligible, the discrepancy between observed revenues increases significantly from that point onward. Specifically, from April 2020 the pandemic impact becomes increasingly pronounced, and only starting to slow down towards the end of the year. Notably, a sudden drop in observed revenue for September 2020 coincides with a major decline in the number of active listings in that month. Even if the difference between observed and counterfactual revenues is relatively small, a couple of notes must be made. In comparison to 2019, our counterfactual estimation is generally more conservative, except for November and December. On the one hand, this indicates that we are not overestimating the effect of the shock. On the other, our machine learning counterfactual imputation approach is picking up patterns that we would not be able to capture if we only looked at 2019 as our baseline.

Figure 3.5: Comparison of observed revenues for 2019 and 2020 vs. counterfactual for 2020

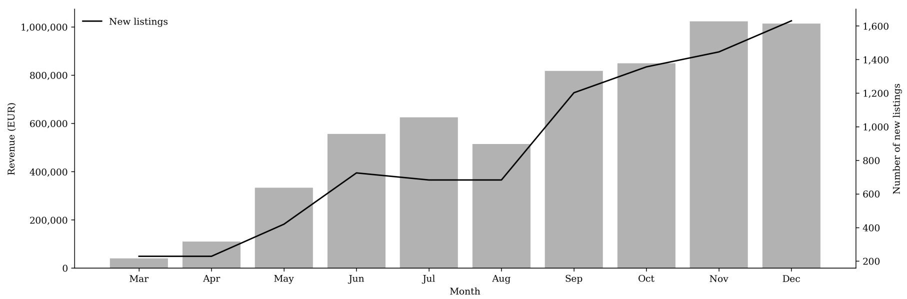


Note: Revenue for active listings during the pre-pandemic period. Counterfactual revenues represent out-sample predictions for the pandemic period.

As discussed, we assume a fixed supply during the pandemic period which means that we do not consider revenue from (i) new listings that entered the market, or (ii) new listings that did not enter but would have entered in case there was no shock. While we cannot measure (ii), we can quantify the general representativeness of

(i) for the Lisbon P2PA market. However, we should expect that their impact is minimal due to *Proposal 204/CM/2019* which took place in 2019 and limited the number of new registrations of P2PAs in Lisbon. In Figure 3.6 we plot the number of new listings and their respective revenues. During the pandemic, a total of 8,602 new properties entered the market, generating a total revenue of 5.9 million EUR, the majority of it coming from the last trimester of the year. However, this figure only represents 2.1% of the counterfactual revenue. This indicates that even if our estimates do not account for new supply, because of the policy intervention in 2019 and the pandemic shock, hosts were deterred from entering the market, and therefore the impact of new supply is negligible. Note that exits are implicitly accounted for within our analysis. By assuming that inactive listings generated zero revenue, we effectively integrate the effects of exits on our estimates.

Figure 3.6: Revenue from new listings in 2020



Note: Number of new listings that were not active during the pre-pandemic period and their respective revenue for the pandemic period.

3.4.4 Robustness

Here we address two key aspects of our analysis to validate our findings. First, we assess our models' stability and whether they are anticipating the pandemic shock by conducting a "placebo" or falsification test. Next, we examine the stability of our model's predictions to understand potential deviations in our estimates.

Placebo Test

Assuming that in February 2020 the Lisbon P2PA market had yet to be significantly impacted by the pandemic, we can treat estimates for this month as a placebo test – akin to the commonly employed ones in synthetic control methods (Abadie et al., 2010). A substantial deviation between observed and counterfactual values for this period — before the pandemic's onset — would imply that our model might inadvertently forecast a shock even when no genuine external disruptions are present. First, in Appendix – 3.8.6, we show that the out-of-sample performance for February 2020 mirrors the average one across the entire pre-pandemic sample. Second, following the specification proposed by Liu et al. (2022), we compute the ATT for this month and verify that it is not statistically different from zero.[15] These results demonstrate that the model's predictions do not prematurely or inaccurately attribute pandemic effects to a time when they should not be present. More specifically, it lends support to Assumptions 1 and 2.

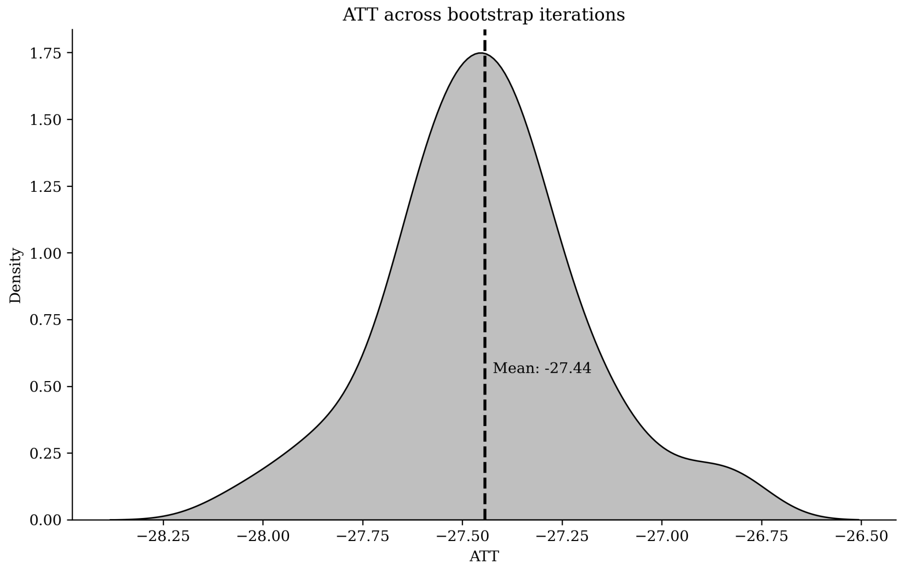
Predictive Uncertainty

When conducting inference with our machine learning counterfactual, it's essential to acknowledge the inherent uncertainty during the imputation phase. The model's prediction accuracy may differ depending on the training data used. To address this, we employ a bootstrap approach, which is elaborated upon in Appendix – 3.9.2. Using this method, we determine the distribution of values for the ATT depicted in Figure 3.7.

The mean effect across bootstrap rounds is -27.44 EUR with a 95% confidence interval (CI) of -27.93 EUR to -29.96 EUR. Our primary estimate of the ATT at -26.92 EUR per day per listing is higher (i.e., more optimistic) than the bootstrap mean, standing at the upper bound of the CI and the right tail of the distribution.

[15] T-statistic of -1.93, with a corresponding p-value of 0.06, of comparing the ATT of -1.65 EUR against a null hypothesis of zero effect.

Figure 3.7: Distribution of ATT across bootstrap iterations



This provides us with additional confidence in the robustness of our primary results, in particular in that we are not underestimating the effects of the pandemic. This not only validates the profound impact but also emphasizes that the observed downturn in revenue, even in the most conservative scenario, is considerably significant.

3.5 Discussion

The COVID-19 pandemic profoundly affected the Lisbon P2PA market, as our study underscores. Like Milone et al. (2023), our study tackles the endogeneity issue between prices and bookings. However, our approach differs in that it does not rely on cross-country variability and extraneous variables. This methodological distinction enables us to provide more precise causal estimates tailored to a specific location. Our findings indicate a marked vulnerability of the market to demand fluctuations. While direct comparisons with other studies are challenging due to variations in the outcome variables, geographical and temporal focus, and identification methods, the magnitude of our observations aligns closely with those of Cheng et al. (2023). They reported a 57.8% decline in global booking activities. However, our findings are more conservative than Chen et al. (2021), who observed an 89.5% income reduction for the Sydney P2PA market in just one month. These losses pose a multitude of challenges not only for P2PAs hosts, who heavily rely on renting out their properties for income, but also for the impacted municipalities, and tourism-related businesses, like restaurants and leisure providers. The additional loss of millions of euros of forsaken taxes and sales revenues should be added to the discussion when reflecting on our results.

Methodologically, we believe that our approach with endogenous covariates addresses an important limitation in the existing counterfactual imputation literature (Borusyak et al., 2023; Souza, 2022), which overlooks the potential endogeneity of covariates, operating under the assumption that these covariates are predetermined or exogenous and thus unaffected by the treatment. This advancement is particularly relevant for dealing with time-varying confounders that can influence the outcome of a shock or an intervention. Furthermore, we demonstrate that, in our context, machine learning algorithms are more effective than linear models in generating accurate counterfactual predictions (i.e., imputations), allowing us to retrieve more precise causal effects.

We believe this is a problem important enough that apply in settings other than P2PA, especially because most current datasets are "big" enough to derive benefits from using machine learning algorithms. Even though we apply it here to the context of a demand shock in the P2PA market, it can easily be extended to several other applications. These include, for instance, labor economics, where generating a counterfactual to estimate the causal effect of economic downturns on job markets might be distorted if one uses post-downturn educational attainment levels as a predictor. These levels might have been influenced by the economic conditions themselves, making them unreliable as post-shock indicators. In healthcare, when trying to discern the causal effect of a new treatment protocol on patient outcomes, using post-treatment medication dosages can be problematic, as these dosages might be adjusted based on the very effects of the new protocol. In environmental economics, estimating the causal impact of new pollution controls on air quality could be skewed if relying on post-control industrial growth rates, since these rates might have changed in direct response to the new regulations. Lastly, in finance, when analyzing the causal effects of regulatory changes on market behavior, using investor sentiment levels recorded after the change might lead to biased conclusions, especially if the sentiment has been shaped by the immediate aftermath of those very regulatory changes. These scenarios underscore the intricate challenges of using post-shock values of endogenous predictors when constructing counterfactual models, highlighting the necessity of our methodological approach in addressing these biases. Furthermore, even if not considered as the main specification, our approach can be used as a complementary analysis or as a robustness test to the main results, namely in problems where there is, in fact, a control group (unlike our setting).

We encourage future research to delve deeper into the potential of this machine learning counterfactual imputation methodology. Exploring its application across diverse contexts, policy types, shocks, and econometric challenges would not only validate its robustness but also refine its capabilities. A particular direction would be the adaptation of our approach to accommodate multiple endogenous covariates. Such advancements would significantly augment the toolset available to applied researchers and practitioners, especially those aiming to pinpoint causal effects in scenarios lacking a definitive control group.

3.6 Conclusion

This study extends the machine learning counterfactual imputation methodology to evaluate the causal effect of the COVID-19 pandemic on the Lisbon P2PA market. We overcome the unaddressed problem in the literature of having an endogenous covariate to the shock. We define a five-step approach that involves estimating an auxiliary model for the endogenous covariate (price) and then using the predicted variable, as well as its out-of-sample residual, as a covariate in the main model (revenue). This approach is similar in spirit to an instrumental variable approach, while leveraging machine learning algorithms to explore complex interactions between our covariates and generate our counterfactual for 2020 without the shock.

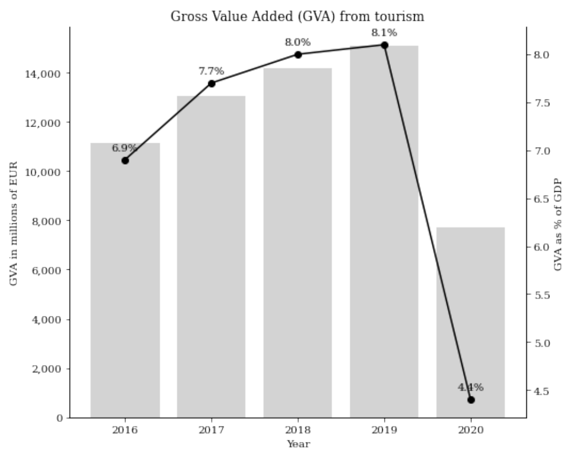
Our results show that the pandemic generated a major demand shock to the Lisbon P2PA market, with the market contracting 43.23%. This represented a total revenue loss of around 121 million EUR during the pandemic period or an average daily loss for each listing of 26.43 EUR. These figures not only underscore the magnitude of the shock but also demonstrate the lack of resilience of the P2PAs to demand shocks.

Appendix

3.7 Data Appendix

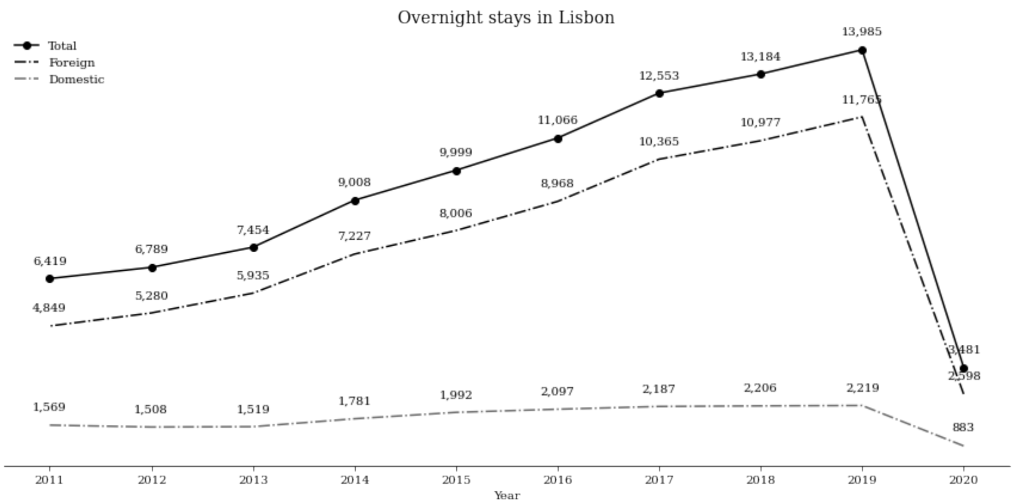
3.7.1 Tourism Data

Figure 3.8



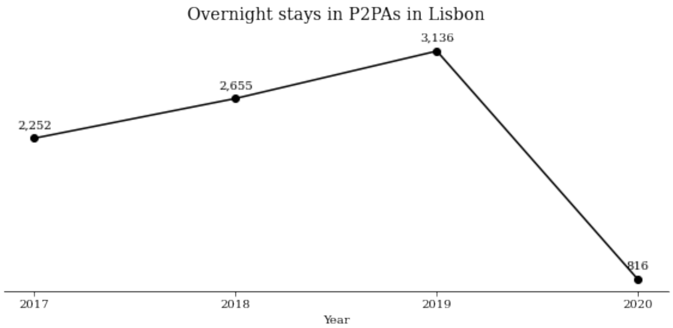
Source: Statistics Portugal

Figure 3.9



Source: Statistics Portugal

Figure 3.10



Source: Statistics Portugal

3.7.2 Our Data

The data used in this study was obtained from Inside Airbnb, a website that collects publicly available data from Airbnb listings around the world.[16] Specifically, we collected monthly scrapes of the website, for the district of Lisbon, Portugal, for the full years of 2019 and 2020. We then focus on the Lisbon Municipality. The scrapes contained essentially two different files: *listings* and *calendar*. Whereas the *listings* file contains the specific characteristics of each listing (ex. number of bedrooms) on the Airbnb platform, *calendar* registers the daily price (on the day of the scrape), availability (true or false), and other booking details for the listing's next 365 days. Both *listings* and *calendar* are monthly snapshots. Even though *calendar* has data for one year into the future, we only use 30 days of it and disregard the remaining dates, because we collected monthly scrapes from the website.[17] Thus, with each monthly scrape, we get an update on the listings' characteristics through *listings*, as well as on the listings' daily prices and availability for the next 30 days with *calendar*. Because listings can be deleted from Airbnb, our data reflects a monthly

[16] http://insideairbnb.com/

[17] The only exception was July of 2020 for which there was no available data from Inside Airbnb. To accommodate for this, instead of considering only 30 days into the future from *calendar*, we use 60 days of data from June's 2020 scrape to complete the missing month of July. Additionally, there may be discrepancies in the number of observations for certain dates in some calendar files due to variations in the starting date of a scrape, which can change depending on when the data was collected, and the lag time between scrapes. Therefore, to account for these variations, for certain scrapes that showcase a sudden drop in the number of observations, we use more than 30 days of data from the preceding scrape.

snapshot of only the listings available on the platform. In other words, we account for listings that exit Airbnb. Finally, we assume that if a property is not available it is because it is booked.[18]

Due to our empirical strategy, discussed in detail in Section 3.3 of the main text, we are only interested in listings' features that are unaffected by (i.e., uncorrelated with) the COVID-19 pandemic shock. For this reason, even though *listings* contemplates more than 100 variables, we are only interested in 18 of them (not accounting for listing identification).[19] We use most of the available features from *calendar*, namely the *price* and *available* variables, which allow us to compute our outcome variable of interest, revenue. Note that the availability variable can then be converted into the number of bookings if aggregated at a listing level, another frequency (ex. monthly), or both (ex. the monthly number of bookings for a specific listing). The complete list of variables used in our analysis and their respective description can be found in Table 3.1. Furthermore, in Table 3.2 we also present descriptive statistics for the selected discrete covariates for both the training and counterfactual periods.[20] Notice that the difference between the means is negligible for all covariates attesting to the fact that most of them are static and unaffected by the pandemic shock.

Finally, we complement our data with the COVID-19: Stringency Index, from Hale et al. (2021), which records the strictness of government policies in a given country, to analyze how the effect of the pandemic relates to local stringency measures adopted.[21] The index is a composite measure of the average of nine of the response metrics, which measure events from workplace closures to international travel controls. A higher value translates into stricter containment measures.

[18]Original data assumptions: http://insideairbnb.com/data-assumptions/

[19]The full list of variables for both listings and calendar files can be here.

[20]The values presented here underwent the process of cleaning and preprocessing described in Subsection 3.7.3

[21]https://ourworldindata.org/covid-stringency-index/

Table 3.1: Variable description

listings

Name	Description	Type
id	Airbnb's unique identifier for the listing	discrete
host_identity_verified	Whether the host has verified his/her identity on the platform	dummy
host_is_superhost	Indicator of host being a "superhost" (more experienced hosts with better reviews)	dummy
host_has_profile_pic	Indicates if the host has a profile picture	dummy
calculated_host_listings_count	Number of listings the host has in the current city scrape	discrete
latitude	Uses the World Geodetic System (WGS84) for latitude	continuous
longitude	Uses the WGS84 projection for longitude	continuous
neighbourhood_cleansed	Neighbourhood as geocoded using latitude and longitude; corresponds to the district's civil parishes for Lisbon	categorical
accommodates	Maximum capacity of the listing	discrete
bathrooms	Number of bathrooms	discrete
bedrooms	Number of bedrooms	discrete
beds	Number of beds	discrete
instant_bookable	Indicator if the guest can automatically book the listing	dummy
require_guest_profile_picture	Requirement for guests to have a profile photo before booking	dummy
require_guest_phone_verification	Requirement for guests to have a phone number before booking	dummy
bed_type	Type of bed	nominal
property_type	Self-selected property type by the host	nominal
room_type	Can be one of: "Entire home/apt", "Private room", "Shared room", "Hotel room"	nominal
host_since	Date when the host/user account was created	date

calendar

Name	Description	Type
listing_id	Airbnb's unique identifier for the listing	discrete
date	Date on the listing's calendar	date
available	Indicates if the date is available for booking	dummy
price	Price listed for the day	continuous

Table 3.2: Metrics for pre-pandemic and pandemic periods (Non-Binary Variables Only)

	Pre-pandemic				**Pandemic**					
Variable	**Mean**	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**	**Mean diff.**	**P-val**
accommodates	3.80	0.04	1	32	3.77	0.02	1	32	-0.03	0.00
bathrooms	1.31	0.01	1	4	1.30	0.00	1	4	-0.02	0.00
bedrooms	1.52	0.01	1	5	1.51	0.00	1	5	0.00	0.00
beds	2.27	0.08	1	9	2.25	0.03	1	9	-0.02	0.00
calculated_host_listings_count	7.97	0.32	1	25	8.05	0.35	1	25	0.08	0.00

Note: The mean and standard deviations were computed at the individual listing level and then averaged, for each of the pre-pandemic and pandemic samples. P-values for the difference in means.

3.7.3 Cleaning and Preprocessing

Missing Values

Some of the variables presented had missing values. We take advantage of the panel structure of our data and first attempt to fill the missing values with information available from previous or future scrapes, as our first step. We follow this approach for the following variables:

host_since, host_is_superhost, host_has_profile_pic, host_identity_verified, instant_bookable, neighbourhood_cleansed, property_type, accommodates, bathrooms, bedrooms, beds, bed_type, require_guest_profile_picture, require_guest_phone_verification

However, for some of these variables, missing information persisted. Thus, in a second step, we imputed their missing values according to their data types:

- Dummy variables: we assumed that if they were missing it was because the condition was not verified (ex. the host did not present a profile picture):

  host_is_superhost, host_has_profile_pic, host_identity_verified, instant_bookable, require_guest_profile_picture, require_guest_phone_verification

- Nominal or categorical variables: even though they might change from one scrape to another (i.e., from month to month), they are typically static variables. Furthermore, these variables typically have a dominant class or value, which is the most popular option in a city (ex. two beds). For these reasons, we considered their most frequent value (i.e., their mode) as the imputation strategy.

bedrooms, bathrooms, beds, property_type, accommodates, bed_type

For the other variables, in which missing data remained after the "first step" we adopted a specific approach which was allowed by the structure of our data:

- *host_since*: imputed with the first date on which the listing shows up in *calendar*.

- *neighbourhood_cleansed*: using the latitude and longitude coordinates from each listing and matching them to their respective civil parish using both public shapefiles and the polygons provided by Inside Airbnb (file called *neighbourhoods*).

It is noteworthy that *price* also has some missing values. Even though they are negligible, representing less than 0.01%, we drop these observations.

Outliers

Some variables presented extreme values that were either inconsistent with their respective variable definition or deviated extremely from their historical values or peers (i.e., properties with similar characteristics). For this reason, we conducted the following procedures:

- *calculated_host_listing_counts*: we assume no host should have less than zero properties; in which case we assume the host has one property. Furthermore, we cap the maximum number of hosts' listings at 25 listings.

- *beds*, *bedrooms*, and *bathrooms*: capped at 9, 5, and 4 respectively, as these represented the 98% percentiles of the respective distributions.

- *price*: first, we winsorized individual listing prices by replacing values exceeding 1.5 times the interquartile range (IQR) above the third quartile, aiming to mitigate excessive price fluctuations for each listing. Subsequently, we winsorized the entire dataset of listing prices, capping the top values at the 98$^{\text{th}}$ percentile. This should address potential data entry errors, such as 999,999 EUR, and outliers. Notably, these percentiles and corresponding price caps were computed and applied separately for the pre-pandemic and pandemic periods for two primary reasons: (i) to prevent data leakage during model selection and training, and (ii) to preserve the inherent price variability differences between the two periods.

Additional Variables

The variable *host_since*, which refers to a date, was converted into a count variable called *host_since_days*, which was used instead. This variable reflects the difference, in the number of days, between the *host_since* date and the booking's current date. The longer the host has been using the platform (i.e., the older the *host_since* date) the higher the *host_since_days*. This should reflect the host's experience with the platform.

The variable *property_type* is a long list of different types of properties, some of which are very rare to observe (ex. lighthouse). For this reason, we map each value to a smaller set of categories and then use them as dummies: *apartment*, *hotel*, *house*, *outdoor*, *special*, and *other* property types.

In addition to the variables presented above, we also consider the following time fixed effects, which we extract from the *calendar* variable *date*: day-of-year, day-of-the-week, month, and a dummy variable for national holidays.[22]

[22]Whether the dates are one of the following: 1st of January (New Year's Day), 25th of April

3.8 Model Selection, Tunning and Training

3.8.1 Time-series Cross-validation

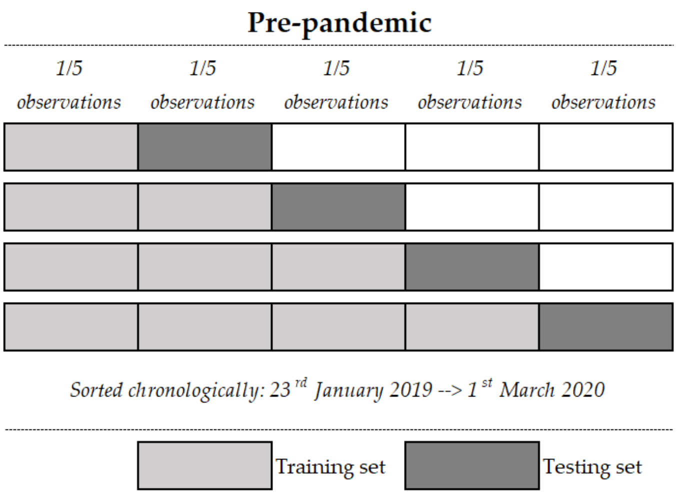
In panel data contexts, there are two possible cross-validation strategies: "vertical" and "horizontal". Whereas the first one stratifies the sample based on units, such as listings in our case, the latter implies splitting across time. We adopt the horizontal approach (i.e. splitting across time) for two reasons. Firstly, our primary objective is to impute a machine learning counterfactual for previously existing listings (fixed supply) for an unseen set of dates. Secondly, vertical cross-validation can lead to overfitting for listings present in the training set. Such overfitting compromises the out-of-sample performance of our machine learning models and it would therefore introduce bias in our estimates.

The "horizontal" or time-series cross-validation (Hyndman and Athanasopoulos, 2021), which we apply to select and tune our models, involves dividing our panel data into a sequence of training and test sets, where the test sets consist of observations that occur after the training set. In other words, each training set is constructed using only observations that occurred before the observation in the corresponding test set. Thus, the earliest observations are not included in the test sets, and these are created by moving forward in time, with each new test set containing the next observations in the time series. By doing this, no future observations can influence the construction of the forecast. To ensure that the training set is large enough for accurate predictions, we split our data into five folds according to the density of our sample, while still preserving the time series component of our data (i.e., listings are sorted chronologically). This means that every listing might show more than once in each fold or training set. The model performance is then calculated by averaging the results over all the test sets, which provides an estimate of how well

(Freedom Day), 1st of May (Labor Day), 10th of June (Portugal Day), 15th of August (Assumption Day), 5th of October (Republic Day), 1st of November (All Saints' Day), 1st of December (Restoration of Independence), 8th of December (Immaculate Conception), and 25th of December (Christmas Day).

the model is likely to perform on new, unseen data (i.e., out-of-sample and future data). Figure 3.1 illustrates this approach.

Figure 3.1: Time-series cross-validation



3.8.2 Model Selection

As detailed in Section 3.3, to impute our machine learning counterfactual we require at least two different models: one auxiliary for forecasting prices; and a main one to predict revenues. Both models are regression ones. The first requirement, which follows, is selecting the best-performing algorithm for each of our regression problems.

We train and evaluate the most commonly used machine learning algorithms namely random forest, XGBoost, LASSO, and multi-layer perceptron (MLP) or neural network. We complement these algorithms with the traditional linear regressions and other baseline algorithms to compare our results with, including decision trees and simple heuristic algorithms (mean and median). For the machine learning algorithms, we use the default hyperparameter settings. While we acknowledge that this may penalize the performance of algorithms that are known to require detailed and elaborate hyperparameter tuning, such as neural networks, it significantly enhances the replicability and accessibility of our work. This approach not only allows other researchers to replicate our findings without extensive computational resources but also ensures a fair comparison of all models in their default state, highlighting their immediate applicability to the specific dataset. In the next section, we then optimize hyperparameters through cross-validation for the best out-of-the-box model. This strategy optimizes the use of computational resources, ensuring the final model fully realizes its potential within our study's context. We use the algorithm versions of the Python libraries Scikit-Learn and Microsoft's LightGBM (a more efficient application of the XGBoost algorithm) for our chapter, but their R, Stata, and other programs' counterparts should display the same easiness of access and performance.[23] In total, we consider six different regression algorithms, plus two simple heuristics.

To select our best-performing algorithm, we apply the time-series cross-validation approach described above to our pre-pandemic data. This allows us to assess the out-of-sample prediction performance of our algorithms. By using only data from our pre-pandemic period to select and then train our best model, we guarantee that our model is not biased from the pandemic period, thereby allowing for the imputation of the machine learning counterfactual. Finally, we the different variables described in Subsections 3.7.2 and 3.7.3 as predictors in our models, as well as their implicit interactions for the decision trees, random forests, XGBoost, and MLP models.

To evaluate the regression models' overall performance, we use the following commonly used metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R-squared or R^2). While RMSE and MAE indicate the accuracy of the model's predictions, with lower values indicating better performance, R-squared measures the proportion of variance in the target variable that is explained by the model, with higher values indicating better fit.

As a result of our five-fold time-series cross-validation strategy, we end up with results for the previous metrics for four test sets (four out-of-sample folds). We

[23]The full documentation for the models used, including the algorithms' default settings and hyperparameters, can be found at https://scikit-learn.org/ and https://lightgbm.readthedocs.io/.

determine our best model by choosing the one that presents the highest overall mean for these metrics over the four test sets.

Table 3.1: Model performance metrics – Price

Model	**MAE**	**RMSE**	**R^2**
Random Forest	15.18	29.51	0.72
Decision Tree	17.55	35.23	0.60
LightGBM	24.46	38.10	0.55
MLP Neural Network	29.51	42.25	0.44
Linear Regression	33.53	46.92	0.31
Lasso Regression	34.60	48.37	0.27
Dummy (Mean)	39.48	57.69	-0.03
Dummy (Median)	38.22	59.80	-0.11

Note: Mean performance across time-series cross-validation folds. The use of out-of-sample residuals may result in negative estimated R^2.

Table 3.2: Model performance metrics – Revenue

Model	**MAE**	**RMSE**	**R^2**
Random Forest	31.82	45.95	0.39
Decision Tree	32.42	61.35	-0.10
LGBM	35.15	45.69	0.39
Linear Regression	36.70	49.82	0.28
Lasso Regression	37.19	49.82	0.28
MLP Neural Network	38.15	52.98	0.17
Dummy (mean)	44.72	60.20	-0.05
Dummy (median)	44.79	61.76	-0.11

Note: Mean performance across time-series cross-validation folds. The use of out-of-sample residuals may result in negative estimated R^2.

In Tables 3.1 and 3.2 we show the results of our model selection process, specifically the average out-of-sample metrics across the four test folds of our time cross-validation approach. It can be noted that, in line with the typical results obtained in the machine learning counterfactual imputation literature, all the machine learning algorithms, except the LASSO, outperform the linear regression model. The algorithm that showcases the overall best performance across the different metrics considered is the random forest. For this reason, we use it as our main algorithm for the pricing and revenue models.

3.8.3 Direct vs. Indirect Revenue Model

Given our data structure, which includes variables for *price* and booking status (*available*), there are two alternatives to conduct the machine learning counterfactual imputation for revenues. Both require the previous estimation of an auxiliary pricing model as discussed in Section 3.3. First, estimate a classifier for booking status and multiply the predicted price by the predicted availability (booked or not booked) to obtain forecasted revenue. Second, and the one we adopt, estimates revenue *directly*. Thus, we call the first option "indirect" and the second one "direct" methods, respectively. We base our choice on the out-of-sample performance for our best-performing model, the random forest, which is also the best-performing algorithm for the classification task (random forest classifier) in the "indirect" alternative. The results for the first approach are the ones presented in Table 3.2. For the "indirect" approach the values for MAE, RMSE, and R^2 are 33.12, 58.64, and -0.01, respectively, which are all lower, demonstrating an overall worse performance.[24]

3.8.4 Model Tunning

To improve the out-of-sample performance of our models we conduct hyperparameter tuning following the time-series cross-validation approach. More specifically, for the best-performing algorithm for both price and revenue, the random forest, we try the following hyperparameter combinations to improve the performance of our models:[25]

- Number of trees: 100.
- Maximum number of features considered for splitting each node: (i) all features; (ii) square root of the number of features; and (iii) logarithm base 2 of

[24] As explained above we have used the default hyperparameter values.

[25] We limit our selection of hyperparameters to avoid making the tunning process more expensive in terms of computational resources.

the number of features.

- Maximum depth of each decision tree: none, 10, 20.
- Minimum number of samples required to split a node: 2, 10.
- Minimum number of samples required to be at a leaf node: 1, 10.

It should be noted that the random forest algorithm does not estimate hyperparameters. Instead, the researcher selects these before running the algorithm. Following the standard approach, for our best model, the random forest, we try all combinations of these hyperparameters and select the best-performing combination on a certain predetermined criterion. In our instance, we select the configuration that shows the lowest MAE, across the four different test sets of our time-series cross-validation strategy. We use MAE to minimize the mean bias of our models and also because there are a lot of extreme values and outliers for prices in our data.

After conducting hyperparameter tuning for both the price and revenue models, the resulting best hyperparameters were as follows: (i) maximum depth of none, (ii) maximum number of features set to square root, (iii) minimum samples per leaf set to 1, (iv) minimum samples per split set to 2 – and a total of 100 estimators. Furthermore, the metrics for the tuned random forest regression models, which should be compared to the values in Tables 3.1 and 3.2, are as follows:

Table 3.3: Model performance metrics – Tuned price and revenue

Model	**MAE**	**RMSE**	**R^2**
Price	14.23	24.91	0.80
Revenue	30.21	44.34	0.43

Note: Mean performance across time-series cross-validation folds.

3.8.5 Model Training

Having selected the best hyperparameters for the price and revenue models, we then train them on the pre-pandemic data and deploy them to the pandemic period, thereby conducting machine learning counterfactual imputation. Reminder to the reader: the reason why we can deploy our model to the pandemic data is because (i) we are only using predictors that should not be affected by treatment (i.e., the pandemic), and (ii) we are training our models only on pre-pandemic data. Note that contrary to when we perform model selection, we train the models on the full pre-treatment data without conducting cross-validation. The time-series cross-validation uses the last fold as a test set, which means that its observations are not included in training the model, which is not the case when we finally train the model – as we include the last test set in the data for training the models. Finally, in considering fixed supply (no new listings) for the pandemic period, we make certain that the number of observations in both the training and imputation phases remain consistent.

3.8.6 Placebo Test

We define March 1, 2020 as the onset of the pandemic. As such, February 2020 should present as a normative month, consistent with the patterns for the AML, depicted in Fig 3.1. This approach resembles a placebo test, adapted to a machine learning context. By training our model on all available pre-pandemic data, excluding only February 2020, and then using this particular month as a testing sample, we anticipate its out-of-sample performance to mirror that of the time-series cross-validation average in Table 3.3. This expectation is corroborated by the findings in Table 3.4.

Interestingly, the performance metrics for February 2020 not only align with but even surpass the average training outcomes. This underscores our model's adeptness in handling unseen data points, bolstering our confidence in its capability to

Table 3.4: Model performance metrics – Placebo for February 2020 (revenue)

Model	**MAE**	**RMSE**	**R2**
Revenue	24.29	34.13	0.56

Note: Out-of-sample performance for the tuned model trained on the pre-pandemic period, except for February 2020, and tested on that month.

accurately capture the pandemic's effect on revenues.

3.8.7 Feature Importance

The feature-importance plots for each model are displayed in Figures 3.2 and 3.3. These plots tell us which features in our data are the most important for predicting our outcome variables (prices and revenues) in training our models.[26] As expected, listings' capacity – as given by the number of beds, bathrooms, or the number of guests it can accommodate – is an important feature in determining price and revenues. Another important takeaway is that location indeed matters, with longitude and latitude being important predictors for both price and revenue. Furthermore, host experience (as given by *host_since_days*) seems to be an important factor. Finally, and importantly, price – as given by the forecasted price from the auxiliary pricing model (*P_hyp*) – is the most important predictor for the main model. Not only that but the average prediction residuals for each listing from the auxiliary model (*residual*) is the fourth most important predictor in our main model which is consistent with endogeneity of the price variable. This indicates that our proposed mechanism, discussed in Subsection 3.3.2, to correct possible bias from the first step also improves the predictive ability of our main model.

To attest to the importance of the predicted price from the auxiliary model to the main model, we evaluate the out-of-sample performance of the random forest model

[26]Specifically, in a random forest, feature importance is obtained by calculating the importance of each feature in predicting the target variable. This is done by measuring the decrease in the model's performance when a particular feature is removed, or its values are randomly shuffled. The higher the decrease in performance, the more important the feature is considered to be.

Figure 3.2: Feature importances – Price

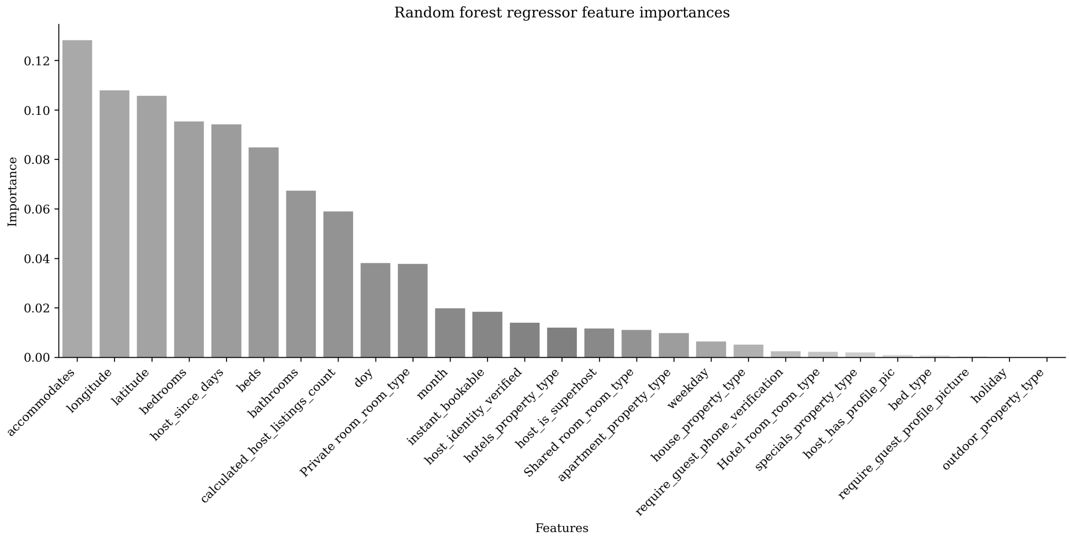
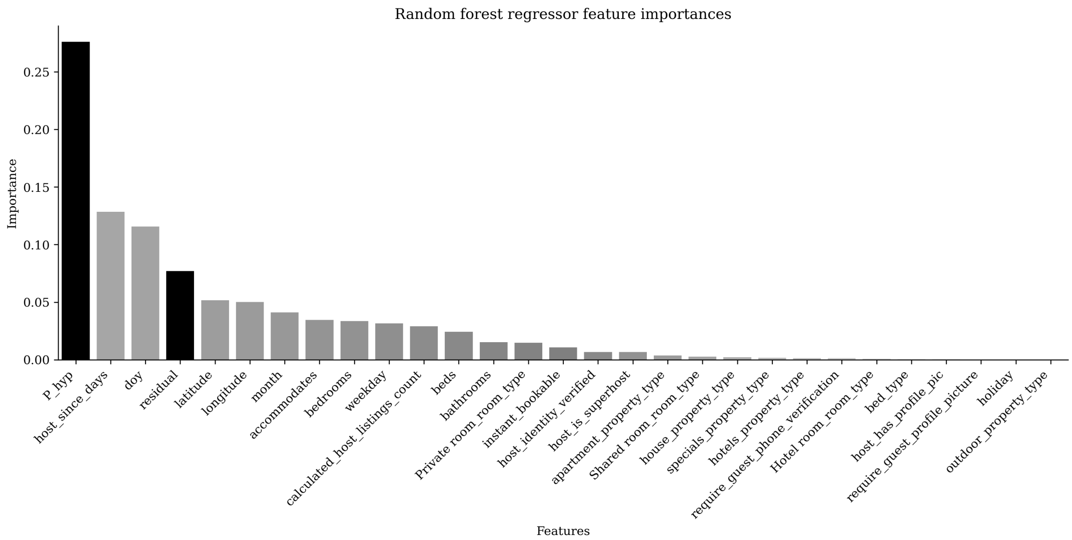


Figure 3.3: Feature importances – Revenue



without using it as a predictor. We do the same for *residual*. Table 3.5 shows that, even without hyperparameter tuning, overall, and as expected, including predicted prices, drastically improves the main model performance. Similarly, adding the predictive residual helps to improve the out-of-sample metrics. These metrics should be compared to the random forest values in Table 3.2, our baseline untuned model. Moreover, adding *residual* as a predictor also reduces bias: we train the (untuned) revenue model without it and confirm the average out-of-sample prediction residual increases – from -0.07 to -0.11.

Table 3.5: Model performance metrics – Revenue without imputed price or with price residuals

Model	**MAE**	**RMSE**	**R2**
With price residual	31.34	45.13	0.41
No imputed price	33.29	51.71	0.22

Note: Mean performance across time-series cross-validation folds.

3.8.8 Prediction Residuals

By using time-series cross-validation, we are able to collect and evaluate the models' out-of-sample residuals. This is important because it is an additional measure of how well the models generalize to unseen data. Furthermore, Souza (2022) shows they should approximate the true counterfactual residuals well. Accordingly, these out-of-sample residuals should be close to zero. This should attest to the model's stability over time (i.e., for new dates), and the avoidance of anticipatory effects, thereby lending support to Assumptions 1 and 2. Finally, it should avoid concerns of potential bias introduced by the machine learning algorithms when imputing the endogenous covariate, as discussed in Subsection 3.3.2.

We plot in Figures 3.4 and 3.6 the out-of-sample and in-sample residuals. These represent the difference between predicted and observed values for the multiple rounds of time-series cross-validation. In Figures 3.5 and 3.7 we present the predic-

Figure 3.4: Residuals – Price

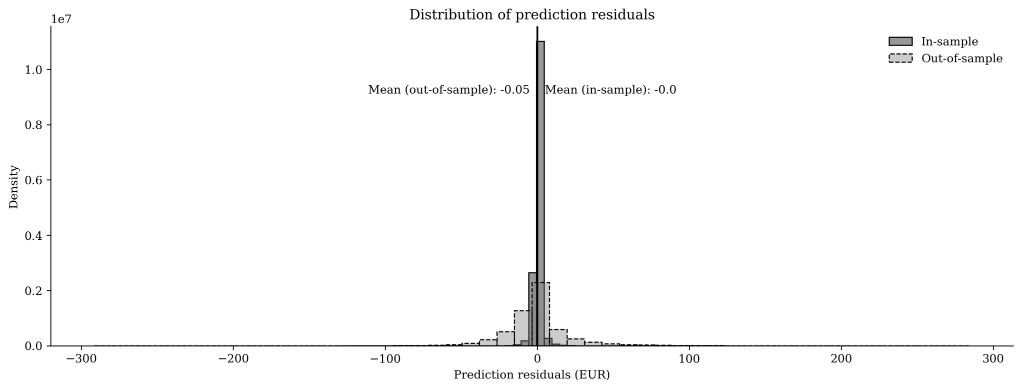


Figure 3.5: Pandemic residuals – Price

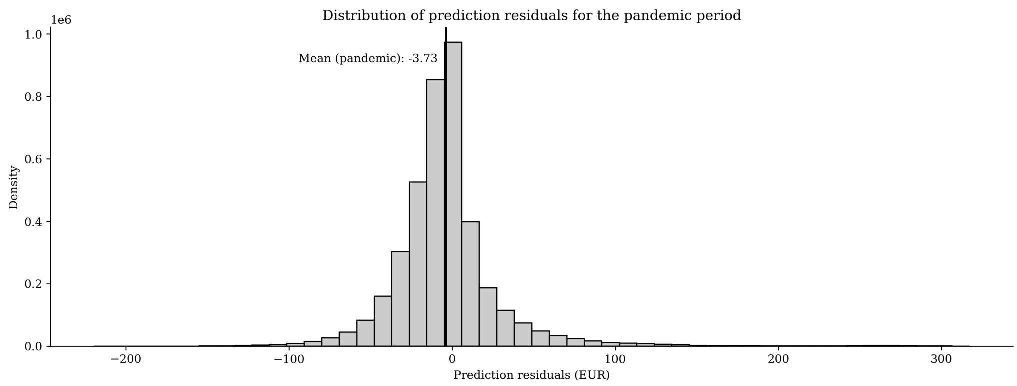


Figure 3.6: Residuals – Revenue

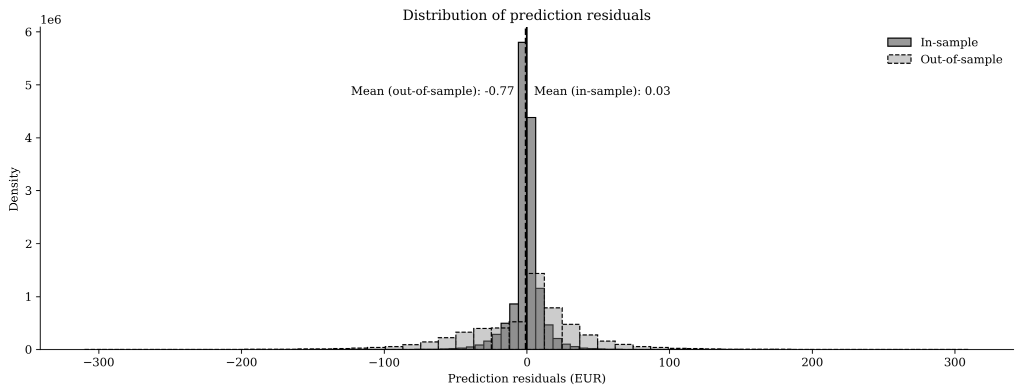
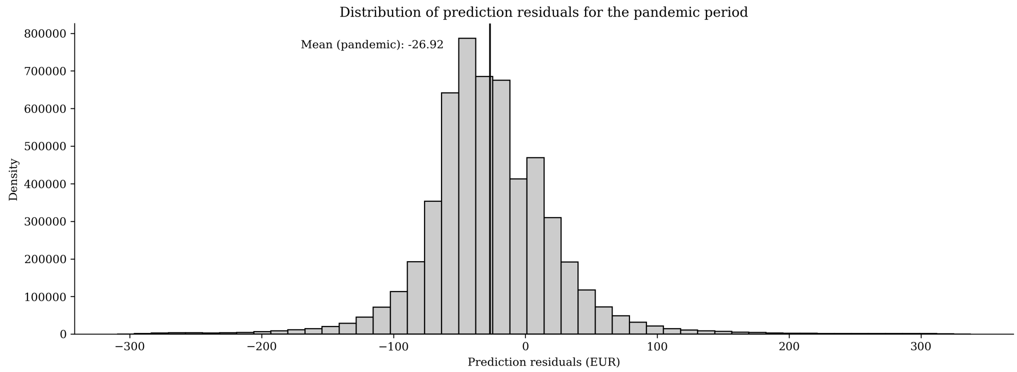


Figure 3.7: Pandemic residuals – Revenue



tion residuals for the pandemic period, which show the difference between observed and (directly) predicted values – that is, the random forest regressor trained on the entire pre-pandemic data and deployed to the pandemic data. There are several important remarks to make about these figures. Firstly, similarly to Souza (2022) we verify that in-sample residuals are more concentrated than out-of-sample (cross-validated) ones. Secondly, and most importantly, the out-of-sample residuals for both prices and revenues resemble a normal distribution centered around zero, indicating that the models have a negligible bias.[27] Thirdly, the distributions for the pandemic period are right skewed, with fatter tails than the out-of-sample residuals, and not centered around zero. This suggests that there was indeed a major shift in the distribution, which is unlikely to represent systematic biases in our models. The fact that both price and revenue pandemic residuals distributions shifted to the left and are centered around negative values, suggests significant price and revenue reductions as a consequence of the pandemic.

[27]The average percentage bias for prices and revenues are -0.06% and 1.52%, respectively.

3.9 Results Appendix

3.9.1 Correlation with Pandemic Progression

Table 3.1 uses stringency index and the total number of COVID-19 cases for Portugal for 2020 and computes the correlation between these metrics and the COVID-19 effect, given by the average revenue loss per day, for the Lisbon area. As expected, there is a clear negative correlation between average daily loss and the stringency index and a positive one for the number of cases. The correlation is stronger for the stringency measure, suggesting that the most important deterrent of demand is in fact mobility constraints. Ultimately, this reinforces our case that the effects we are picking up in our estimates are indeed the effect of the pandemic and not systematic errors in our estimation strategy.

Table 3.1: Correlation between daily revenue loss and the stringency index and total cases

	Correlation	**p-value**
Stringency index	-0.28	0.00
Number of cases	0.17	0.00

3.9.2 Boostrap

When conducting inference with machine learning estimates, it is vital to acknowledge the inherent uncertainty in the predictive phase. The performance of the model's predictions can exhibit variations based on the sample it is trained on. To address this variability, we follow a conservative bootstrap approach to approximate the standard errors of our parameters of interest, as suggested by Souza (2022). This method captures the variability due to differing potential training samples, offering a robust estimate of uncertainty. Our bootstrap algorithm is as follows:

1. **Initialization**:

- Let N represent the total number of unique listings in the pre-pandemic sample.
- Use b to denote each bootstrap iteration, ranging from 1 to B.

2. **Sample creation**:

   - Draw a vector of N_b listings, denoted as $i_1, i_2, \ldots, i_{N_b}$. These listings are chosen randomly from the N available ones in the pre-pandemic period.
   - Create a bootstrap sample by employing stratified random sampling (stratified by listing) with replacement, ensuring that the size of the bootstrap sample, N_b, approximates N.

3. **Machine learning counterfactual imputation with endogenous covariates**: follow steps 1, 3, and 4 of our five-step approach in Subsection 3.3.2, using the bootstrapped sample, and obtain a forecast for counterfactual prices and revenues:[28]

4. **Iteration**: repeat steps 2 to 3 for a total of B bootstrap iterations.

5. **Standard error calculation**: compute the bootstrapped standard error, as given by:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(\beta_b - \hat{\beta})^2}{B}}$$

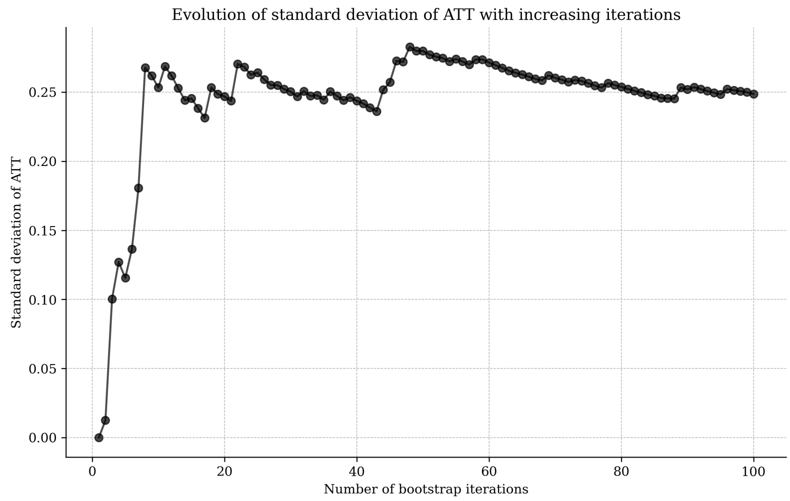
6. **Confidence interval (CI) computation**: determine confidence bounds for the value of interest – either prices, revenues, or the ATT. The lower bound is $\beta^- = \beta - 1.96 \times \sigma$ and the upper bound is $\beta^+ = \beta + 1.96 \times \sigma$.

Figure 3.1 presents the fluctuations in the standard deviations of the machine learning estimates for the ATT, across iterations. A significant degree of volatility is

[28]We assume fixed residuals for step 2 because these are a product of time-series cross-validation, which already incorporates sample variability.

evident during the initial 40 iterations, suggesting that the model's estimates were less consistent during this early phase. By the time we reach 80 iterations, this inconsistency appears to stabilize, with the values fluctuating only marginally. As we approach the 100$^{\text{th}}$ iteration, the oscillation in standard deviations stabilizes at 0.25.

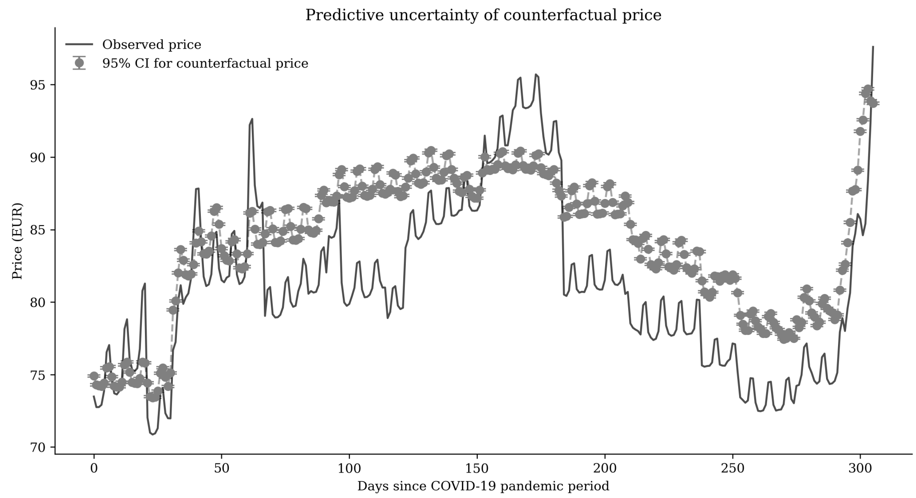
Figure 3.1: Stability of bootstrapped standard deviations



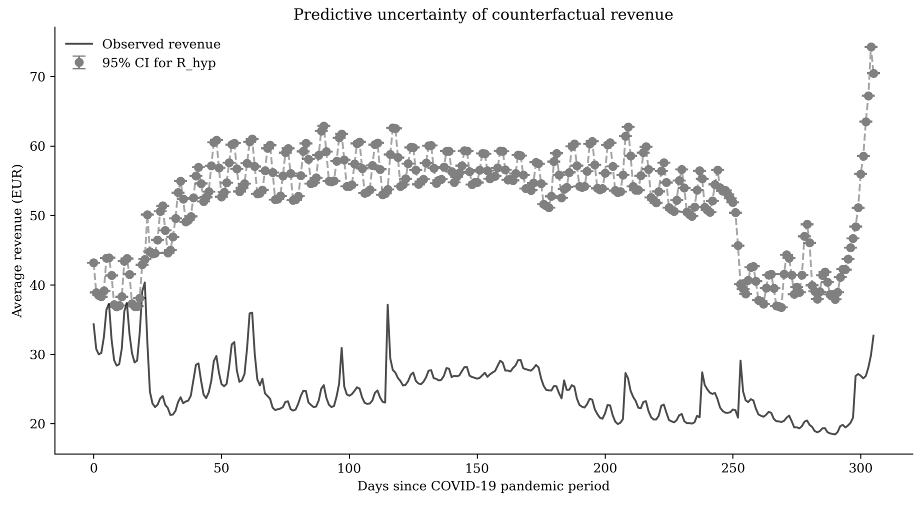
Figures 3.2a and 3.2b represent the distributions of prices and revenues, respectively, for each day as a result of our bootstrap approach. These figures resemble Figures 3.2 and 3.4b presented before. This is because our CIs are notably narrow. This is a result of having more than 20 thousand listings for each date and the fact that the random forest already incorporates bootstrapping as a default – which makes it robust and less prone to overfitting. This attests to our models' stability and our counterfactual imputations being resilient to variations in sample selection.

Figure 3.2: Confidence intervals

(a) Price



(b) Revenue



Conclusion

This thesis has surveyed the landscape of Portugal's real estate sector, unveiling the critical role of data, the impact of STRs, and the effects of global pandemics on urban housing markets. My research underscores the necessity for enhanced data to inform evidence-based policymaking. The analysis of short-term rental dynamics, through the lens of Airbnb's presence in Lisbon, reveals the relationship between such platforms and housing supply, highlighting both opportunities and challenges for urban planning and housing affordability. Furthermore, the application of machine learning for counterfactual imputation, especially incorporating endogenous covariates, improves the derivation of relevant insights into the economic repercussions of shocks like the COVID-19 pandemic on STRs and akin markets.

My findings underscore the value of comprehensive, integrated data and advanced analytical methods in understanding the complexities of real estate markets. They not only contribute to academic discourse but also offer practical guidance for policymakers, stakeholders, and investors aiming to better navigate the real estate landscape. The insights derived from this research provide a foundation for more adaptive, responsive, and informed real estate policy development. Looking forward, the methodologies and discoveries of this thesis pave the way for further research, encouraging a deeper exploration into the intersections of technology, policy, and real estate dynamics on a global scale.

Bibliography

Abadie, A., Diamond, A., , and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105:493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country who study the effects of terrorism on tourism and foreign direct investment in spain. 113. *American Economic Review*, 93:113 – 132.

Adje, K. D. C., Letaifa, A. B., Haddad, M., and Habachi, O. (2023). Smart city based on open data: A survey. *IEEE Access*, 11:56726–56748.

Alexopoulos, C., Saxena, S., Loukis, E. N., Rizun, N., Maratsi, M. I., and Ali, M. (2023). Chatgpt application vis-a-vis open government data (ogd): Capabilities, public values, issues and a research agenda. *SSRN Electronic Journal*.

Almagro, M. and Domínguez-Iino, T. (2022). Location sorting and endogenous amenities: Evidence from amsterdam. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 162.

Arinder, M. K., Isett, K. R., Head, B. W., and Vanlandingham, G. (2016). Bridging the divide between evidence and policy in public sector decision making: A practitioner's perspective. *Public Administration Review*, 76:394–398.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116:1716–1730.

Baron, J. (2018). A brief history of evidence-based policy. *Annals of the American Academy of Political and Social Science*, 678:40–50.

Barron, K., Kung, E., and Proserpio, D. (2021). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, 40(1):23–47.

Batalha, M., Gonçalves, D., Peralta, S., and dos Santos, J. P. (2022). The virus that devastated tourism: The impact of covid-19 on the housing market. *Regional Science and Urban Economics*, 95.

Bauer, T. K., Braun, S. T., and Kvasnicka, M. (2017). Nuclear power plant closures and local housing values: Evidence from fukushima and the german housing market. *Journal of Urban Economics*, 99:94–106.

Bekkerman, R., Cohen, M. C., Kung, E., Maiden, J., and Proserpio, D. (2022). The Effect of Short-Term Rentals on Residential Investment. *Marketing Science*, pages 1–16.

Benatia, D. and Gingras, S. (2023). Reaching new lows? the pandemic's consequences for electricity markets. *The Energy Journal*, 44:195–222.

Berends, J., Carrara, W., and Radu, C. (2020). Analytical report 9: The economic benefits of open data.

Blundell, R., Griffith, R., and Reenen, J. V. (1999). Market share, market value and innovation in a panel of british manufacturing firms. *The Review of Economic Studies*, 66:529–554.

Bodenbender, M., Kurzrock, B. M., and Müller, P. M. (2019). Broad application of artificial intelligence for document classification, information extraction and predictive analytics in real estate. *Journal of General Management*, 44:170–179.

Borusyak, K., Xavier, J., and Spiess, J. (2023). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint*.

Bresciani, S., Ferraris, A., Santoro, G., Premazzi, K., Quaglia, R., Yahiaoui, D., and Viglia, G. (2021). The seven lives of airbnb. the role of accommodation types. *Annals of Tourism Research*, 88:103170.

Brueckner, J. K. (1980). A vintage model of urban growth. *Journal of Urban Economics*, 8(3):389–402.

Brueckner, J. K. (2000). Urban Sprawl: Diagnosis and Remedies. *International Regional Science Review*, 23(2):160–171.

Büchler, S., v. Ehrlich, M., and Schöni, O. (2021). The amplifying effect of capitalization rates on housing supply. *Journal of Urban Economics*, 126:103370.

Cantador, I., Viejo-Tardío, J., Cortés-Cediel, M. E., and Bolívar, M. P. R. (2021). A chatbot for searching and exploring open data: Implementation and evaluation in e-government. *DG. O2021: The 22nd Annual International Conference on Digital Government Research*, pages 168–179.

Cao, L., Xu, P., and Shang, W. (2021). A text-based mining approach for real estate policy impact monitoring and analysis. *IEEE International Conference on Big Data, Big Data 2021*, pages 1575–1581.

Cardoso, F. S. (2022). Official statistics: Historical trajectory and contemporary threat related to big data adoption. *ISLA 2022 Proceedings*.

Cerqua, A. and Letta, M. (2022). Local inequalities of the covid-19 crisis. *Regional Science and Urban Economics*, 92.

Cerqua, A., Stefano, R. D., Letta, M., and Miccoli, S. (2021). Local mortality estimates during the covid-19 pandemic in italy. *Journal of Population Economics*, 34:1189–1217.

Chakraborty, A., Knaap, G. J., Nguyen, D., and Shin, J. H. (2009). The Effects of High-density Zoning on Multifamily Housing Construction in the Suburbs of Six US Metropolitan Areas. *Urban Studies*, 47(2):437–451.

Chen, G., Cheng, M., Edwards, D., and Xu, L. (2021). Covid-19 pandemic exposes the vulnerability of the sharing economy: a novel accounting framework. *Journal of Sustainable Tourism*.

Cheng, M., Hu, M., and Lee, A. (2023). A global perspective on the impact of covid-19 on peer-to-peer accommodation: human mobility, case number and lockdown policies. *International Journal of Contemporary Hospitality Management*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68.

Chetty, B. R., Hendren, N., and Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106:855–902.

Clapp, J. M., Bardos, K. S., and Wong, S. K. (2012). Empirical estimation of the option premium for residential redevelopment. *Regional Science and Urban Economics*, 42(1-2):240–256.

Cócola-Gant, A. (2016). Holiday rentals: The new gentrification battlefront. *Sociological Research Online*, 21(3).

Delisle, J. R. and Grissom, T. V. (2020). The big data regime shift in real estate. *Journal of Property Investment & Finance*.

Dueñas, M., Ortiz, V., Riccaboni, M., and Serti, F. (2021). Assessing the impact of covid-19 on trade: a machine learning counterfactual analysis. *arXiv preprint*.

Duro, J. A., Perez-Laborda, A., Turrion-Prats, J., and Fernández-Fernández, M. (2021). Covid-19 and tourism vulnerability. *Tourism Management Perspectives*, 38:100819.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.

Edelman, B. G. and Luca, M. (2014). Digital Discrimination: The Case of Airbnb.com. *Harvard Business School NOM Unit Working Paper*, 14(054).

European Comission (2023). Digital public administration factsheet 2023 portugal.

Fabra, N., Lacuesta, A., and Souza, M. (2022). The implicit cost of carbon abatement during the covid-19 pandemic. *European Economic Review*, 147:104165.

Filippas, A. and Horton, J. J. (2018). The tragedy of your upstairs neighbors: Externalities of home-sharing. *New York University working paper.*

Franco, S. F. and Santos, C. D. (2021). The impact of airbnb on residential property values and rents: Evidence from portugal. *Regional Science and Urban Economics*, 88.

Galster, G. and Lee, K. O. (2021). Introduction to the special issue of the global crisis in housing affordability. *International Journal of Urban Sciences*, 25:1–6.

Garcia-López, M. À., Jofre-Monseny, J., Martínez-Mazza, R., and Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, 119(103278).

Gerwe, O. (2021). The covid-19 pandemic and the accommodation sharing sector: Effects and prospects for recovery. *Technological Forecasting and Social Change*, 167:120733.

Glaeser, E. L. and Gyourko, J. (2003). The Impact of Building Restrictions on Housing Affordability. *FRBNY Economic Policy Review*, 9:21–39.

Glaeser, E. L., Gyourko, J., and Saiz, A. (2008). Housing supply and housing bubbles. *Journal of Urban Economics*, 64(2):198–217.

Glaeser, E. L., Gyourko, J., and Saks, R. E. (2006). Urban growth and housing supply. *Journal of Economic Geography*, 6(1):71–89.

Glaeser, E. L. and Ward, B. A. (2009). The causes and consequences of land use regulation: Evidence from Greater Boston. *Journal of Urban Economics*, 65(3):265–278.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, pages 535–551.

Guardia, F. H. D. L., Grant, S., and Miguel, E. (2020). A framework for open policy analysis. *Science and Public Policy*, 0.

Gyourko, J., Mayer, C., and Sinai, T. (2013). Superstar Cities. *American Economic Journal: Economic Policy*, 5(4):167–99.

Gyódi, K. (2022). Airbnb and hotels during covid-19: different strategies to survive. *International Journal of Culture, Tourism and Hospitality Research*, 16:168–192.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. (2021). A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5:529–538.

Hart, N. and Yohannes, M. (2019). Evidence works: Cases where evidence meaningfully informed policy.

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics*, 86:30–57.

Hilber, C. A. and Vermeulen, W. (2016). The Impact of Supply Constraints on House Prices in England. *The Economic Journal*, 126(591):358–405.

Hjort, J., Moreira, D., Rao, G., and Santini, J. F. (2021). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. *American Economic Review*, 111:1442–1480.

Horn, K. and Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. *Journal of Housing Economics*, 38:14–24.

Hu, L., He, S., Han, Z., Su, S., Weng, M., and Cai, Z. (2018). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy*, 82:657–673.

Hyndman, R. and Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed).

Ihlanfeldt, K. R. (2007). The effect of land use regulation on housing and land prices. *Journal of Urban Economics*, 61(3):420–435.

Jackson, K. (2016). Do land use regulations stifle residential development? Evidence from California cities. *Journal of Urban Economics*, 91:45–56.

Jang, S., Kim, J., Kim, J., and Kim, S. S. (2021). Spatial and experimental analysis of peer-to-peer accommodation consumption during covid-19. *Journal of Destination Marketing & Management*, 20:100563.

Janssen, M. and van den Hoven, J. (2015). Big and open linked data (bold) in government: A challenge to transparency and privacy? *Government Information Quarterly*, 32:363–368.

Joseph, B. K. (2019). *Blockchain for Open Data – Exploring Conceptual Underpinnings and Practice*, volume 31, pages 161–175. Springer.

Kaplan, R. A. and Nadler, M. L. (2015). Airbnb: A Case Study in Occupancy Regulation and Taxation. *University of Chicago Law Review Dialogue*, 82.

Kok, N., Monkkonen, P., and Quigley, J. M. (2014). Land use regulations and the value of land and housing: An intra-metropolitan analysis. *Journal of Urban Economics*, 81:136–148.

Koster, H. R., van Ommeren, J., and Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics*, 124(103356).

Landvoigt, T., Piazzesi, M., and Schneider, M. (2015). The housing market(s) of san diego. *American Economic Review*, 105:1371–1407.

Law, S., Paige, B., and Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10.

Lee, D. (2016). How Airbnb Short-Term Rentals Exacerbate Los Angeles's Affordable Housing Crisis: Analysis and Policy Recommendations. *Harvard Law Policy Review*, 10(1):229 – 253.

Lemieux, V. and Flores, D. (2017). Real estate transaction recording in the blockchain in brazil.

Leoni, F., Carraro, M., McAuliffe, E., and Maffei, S. (2023). Data-centric public services as potential source of policy knowledge. can "design for policy" help? *Transforming Government: People, Process and Policy*, 17:399–411.

Levine, N. (1999). The Effects of Local Growth Controls on Regional Housing Production and Population Redistribution in California. *Urban Studies*, 36(12):2047–2068.

Lieber, R. (2015). New worry for home buyers: A party house next door. *New York Times*.

Liu, L., Wang, Y., and Xu, Y. (2022). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 00:1–17.

Liu, X., Song, Y., Wu, K., Wang, J., Li, D., and Long, Y. (2015). Understanding urban china with open data. *Cities*, 47:53–61.

Loberto, M., Luciani, A., and Pangallo, M. (2021). What do online listings tell us about the housing market?

Mayer, C. J. and Somerville, C. T. (2000). Land use regulation and new construction. *Regional Science and Urban Economics*, 30(6):639–662.

Mcmurren, J., Young, A., and Verhulst, S. (2018). Addressing transaction costs through blockchain and identity in swedish land transfers.

Mian, A., Rao, K., and Sufi, A. (2013). Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*, 128:1687–1726.

Milone, F. L., Gunter, U., and Zekan, B. (2023). The pricing of european airbnb listings during the pandemic: A difference-in-differences approach employing covid-19 response strategies as a continuous treatment. *Tourism Management*, 97:104738.

Munneke, H. J. and Womack, K. S. (2015). Neighborhood renewal: The decision to renovate or tear down. *Regional Science and Urban Economics*, 54:99–115.

Nagaraj, A., Shears, E., and Vaan, M. D. (2020). Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences of the United States of America*, 117:23490–23498.

Nagaraj, A. and Tranchero, M. (2023). How does data access shape science? evidence from the impact of u.s. census's research data centers on economics research. *National Bureau of Economic Research*.

Ooijen, C. V., Ubaldi, B., and Welby, B. (2019). A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance.

Page, M., Hajduk, E., Arriëns, E. N. L., Gianfranco, C., and Brinkhuis, S. (2023). Open data maturity report 2023.

Pendall, R. (2000). Local Land Use Regulation and the Chain of Exclusion. *Journal of the American Planning Association*, 66(2):125–142.

Quigley, J. M. and Rosenthal, L. A. (2005). The Effects of Land-Use Regulation on the Price of Housing: What Do We Know? What Can We Learn? *Cityscape: A Journal of Policy Development and Research*, 8(1):69–137.

Rodrigues, P. M. M. (2022). The real estate market in portugal.

Rohe, W. M. (2017). Tackling the housing affordability crisis. *Housing Policy Debate*, 27:490–494.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Saari, A., Junnila, S., and Vimpari, J. (2022). Blockchain's grand promise for the real estate sector: A systematic review. *Applied Sciences 2022, Vol. 12, Page 11940*, 12:11940.

Sainaghi, R. and Chica-olmo, J. (2022). The effects of location before and during covid-19 impacts on revenue of airbnb listings in milan (italy). *Annals of Tourism Research*, 96:103464.

Saiz, A. (2010). The Geographic Determinants of Housing Supply. *The Quarterly Journal of Economics*, 125(3):1253–1296.

Saiz, A. (2020). Bricks, mortar, and proptech the economics of it in brokerage, space utilization and commercial real estate finance. *Journal of Property Investment & Finance*.

Saiz, A. and Salazar-Miranda, A. (2023). Understanding urban economies, land use, and social dynamics in the city: Big data and measurement.

Saull, A., Baum, A., and Braesemann, F. (2019). Can digital technologies speed up real estate transactions? *Journal of Property Investment & Finance*.

Schuetz, J. (2020). Teardowns, popups, and renovations: How does housing supply change? *Journal of Regional Science*, 60(3):459–480.

Sheppard, S. and Udell, A. (2016). Do Airbnb properties affect house prices? *Williams University working paper*.

Simoes, V. C., Krieger, K., and Melchor, L. (2022). Science for policy in portugal.

Souza, M. (2022). Predictive counterfactuals for treatment effect heterogeneity in event studies with staggered adoption. *SSRN Electronic Journal*.

Splawa-Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472.

Tan, E. (2021). A conceptual model of the use of ai and blockchain for open government data governance in the public sector.

Tene, O. and Polonetsky, J. (2011). Privacy in the age of big data: A time for big decisions. *Stanford Law Review Online*, 64.

Thorson, J. A. (1997). The Effect of Zoning on Housing Construction. *Journal of Housing Economics*, 6(1):81–91.

Treleaven, P., Barnett, J., Knight, A., and Serrano, W. (2021). Real estate data marketplace. *AI and Ethics*, 1:445–462.

Valentin, M. (2021). Regulating short-term rental housing: Evidence from New Orleans. *Real Estate Economics*, 49(1):152–186.

van Dijk, D. W. and Francke, M. K. (2018). Internet search behavior, liquidity and prices in the housing market. *Real Estate Economics*, 46:368–403.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28:3–28.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113:7310–7315.

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., and Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land*, 11:334.

Weinberg, D. H. (2014). Data sources for u.s. housing research, part 1: Public sector data sources. *Cityscape*, 16.

Weinberg, D. H. (2015). Data sources for u.s. housing research, part 2: Private sources, administrative records, and future directions. *Cityscape*, 17.

Wetzstein, S. (2017). The global urban housing affordability crisis. *Urban Studies*, 54:3159–3177.

Wheaton, W. C. (1978). Price-Induced Distortions in Urban Highway Investment. *The Bell Journal of Economics*, 9(2):622.

Wheaton, W. C. (1982). Urban spatial development with durable but replaceable capital. *Journal of Urban Economics*, 12(1):53–67.

Wooldridge, J. M. (2016). Should instrumental variables be used as matching variables? *Research in Economics*, 70:232–237.

World Bank (2021). World development report 2021: Data for better lives.

Xu, M. and Xu, Y. (2021). What happens when Airbnb comes to the neighborhood: The impact of home-sharing on neighborhood investment. *Regional Science and Urban Economics*, 88(103670).

Zervas, G., Proserpio, D., and Byers, J. W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research*, 54(5):687–705.

Zhang, M., Geng, R., Huang, Y., and Ren, S. (2021). Terminator or accelerator? lessons from the peer-to-peer accommodation hosts in china in responses to covid-19. *International Journal of Hospitality Management*, 92:102760.

Zoğal, V., Domènech, A., and Emekli, G. (2020). Stay at (which) home: second homes during and after the covid-19 pandemic. *Journal of Tourism Futures*, 8:125–133.

Òscar Jordà, Schularick, M., and Taylor, A. M. (2015). Leveraged bubbles. *Journal of Monetary Economics*, 76:S1–S20.