Housing Economics and Policy Evaluation in the UK — New Insights from Big Data

YUNLONG HUANG

A thesis submitted in partial fulfilment of the
requirement of University College London (UCL)
for the degree of Doctor of Philosophy

THE BARTLETT SCHOOL OF SUSTAINABLE CONSTRUCTION
UNIVERSITY COLLEGE LONDON

Declaration

I, Yunlong Huang confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Employing big data techniques, I develop several computer programs to construct large micro-level datasets on residential transactions in England, and demonstrate a solution for standardised and comprehensive data collection methods and frameworks, enabling robust analysis and understanding of housing markets. These novel datasets amalgamate information from various open sources. The smallest dataset encompasses information on one-third of the population's residential transactions, while the largest dataset covers over 92% of all transactions recorded by the Land Registry. To the best of my knowledge, these are among the most comprehensive datasets utilised in similar studies, offering unique insights into two aspects of the UK residential market.

The relationship between transaction price (TP) and time on the market (TOM) remains a longstanding puzzle in the field. Despite extensive research into both the effect of TOM on price and the effect of price on TOM, numerous inconsistent findings persist. In Chapter 3, I address two key issues contributing to these inconsistencies: (i) the omission of controlling for overpricing and (ii) the endogeneity arising from the simultaneous relationship between TOM and price. To tackle these issues, I propose a new overpricing measurement and utilise a two-stage least squares (2SLS) method, employing two novel instrumental variables (IVs): the price revision duration and the council tax band (CTB) for TOM and transaction price, respectively. My results reveal a positive and robust relationship between price and TOM, in line with search theory. Furthermore, my results suggest that chain-free sellers, who are not subject to the constraints of selling their current property to proceed with their next steps, set lower initial asking prices and agree to lower transaction prices, all else being equal, which is associated with agency costs.

Using the 2020 Stamp Duty holiday (SDH) in the UK as a quasi-natural experiment, I provide a comprehensive analysis of the tax reduction's effects on transaction and listing volumes, prices, and market liquidity. Theoretically, I develop a Nash bargaining model and demonstrate that the SDH leads to an increase in prices and a greater surplus for sellers. Empirically, I adopt difference-in-differences (DiD) models and find that the SDH resulted in a 53% increase in housing transactions and an average increase of over 2% in transaction prices; additionally, sellers' bargaining power strengthened as the SDH deadline approached. Most of the tax savings from the SDH were passed on to sellers in the form of increased prices, leading to reduced affordability for first-time buyers and home movers replacing their main residence. I also discover evidence that market participants utilised the SDH to relocate away from highly urbanised, polycentric areas during the Covid-19 pandemic. My findings indicate that while an SDH can stimulate market activity during an economic downturn and enable the housing market to adjust to changing conditions rapidly, it may also inadvertently reduce housing affordability.

Impact Statement

The findings presented in this thesis hold significant potential for providing benefits within and beyond academia. For the academic community, the methodology and approach developed in this research may inspire further innovation and creativity in the field of housing market research. By harnessing big data and novel sources of information, researchers can overcome challenges that have long plagued the study of the UK housing market. The datasets generated through this research rank among the most comprehensive utilised in related studies on the UK market, offering a valuable resource for future investigations. Moreover, the thesis sheds light on the long-standing puzzle of the relationship between price and TOM, a conundrum that has remained unresolved for many years. The proposed 2SLS estimation process, employing novel instrumental variables and a newly constructed measure of overpricing, presents a solution to the model identification issues hindering previous studies, thereby providing a more accurate understanding of search theory in the housing market.

Beyond academia, this thesis' findings bear substantial implications for public policy design and public service delivery in the UK. The research emphasises the potential of big data and open data initiatives to inform public policy, particularly concerning property transaction taxes. Examining the effects of the 2020 stamp duty land tax holiday on the residential market yields crucial insights into the trade-offs between policy objectives such as promoting homeownership, stimulating market activity, and preserving housing affordability. The research findings suggest that while a SDH can stimulate market activity during an economic downturn and facilitate rapid housing market adjustments to changing conditions, it may inadvertently reduce housing affordability. This highlights the importance of carefully weighing the impact of tax policy on the housing market and broader society.

The commercial sector also stands to benefit from the insights generated by this thesis. The research offers a valuable resource for real estate agents, property developers, and other industry professionals seeking to understand the dynamics of the UK housing market. By providing a more accurate understanding of the relationship between price and TOM, and the effects of property transaction tax changes, the research findings can inform business decisions and strategies within the real estate industry.

In conclusion, the research presented in this thesis introduces a new approach to studying the UK housing market, leveraging big data and novel sources of information to provide insights into long-standing puzzles in the field. The findings hold the potential to inform public policy, benefit industry professionals, and inspire further research and innovation in the field.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my parents, who have always believed in me and supported me throughout this journey. Their unwavering love, encouragement, and sacrifices have provided the foundation upon which I have built my academic career. I am forever grateful for the values they instilled in me and the opportunities they have given me.

I would like to extend my sincerest appreciation to my supervisor, Prof. Stanimira Milcheva, who has been an exemplary mentor. Her invaluable advice has been instrumental in shaping my work. I am particularly grateful for the freedom she provided me to explore my research interests and develop my own ideas, which has greatly enriched my academic experience.

To my partner, Qi, words cannot express my gratitude for your companionship, love, and patience during this process. Your unwavering support, understanding, and belief in me have been a source of strength and inspiration. I could not have accomplished this without you by my side.

I would also like to acknowledge my friends and fellow PhD colleagues, whose camaraderie has made this journey not only intellectually rewarding but also enjoyable. Your presence has truly enriched my life, and I am proud to have shared this experience with you.

Lastly, I would like to thank everyone who has contributed to my research in any way, directly or indirectly. Your assistance, support, and encouragement have played a crucial role in my achievements, and I am deeply grateful for your involvement in this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

2SLS Two Stages Least Square.

API Application Programming Interface.

CTB Council Tax Band.

DDD Difference-in-Difference-Differences.

DiD Difference-in-Differences.

DLUHC Department for Levelling Up, Housing and Communities.

DOP Degree of Overpricing.

EPC Energy performance certificates.

IV Instrumental Variable.

ML Machine Learning.

OLS Ordinary Least Squares.

PPD Price Paid Data.

RUC Rural-Urban Classification.

SDH 2022 Stamp Duty Holiday.

SDLT Stamp Duty Land Tax.

SEM Simultaneous Equations Model.

TOM Time on Market.

TP Transaction Price.

UPRN Unique Property Reference Number.

VOA Valuation Office Agency.

Chapter 1

Introduction

1.1 Background

Data are crucial for both research and policy development. Over recent decades, economic research has increasingly adopted an empirical approach. Consequently, economists are frequently sought by the media for their expert opinions on contemporary public issues and to provide testimony during policy-making and legislative processes. This shift can be partly attributed to the rapid proliferation of internet connections, which facilitates the compilation of datasets from diverse sources and provides easier access to more cost-effective computational power for performing intricate analyses. The increasing availability of data has led economists to rely on real-world information to supplement and test their theoretical models.

In an analysis of 748 research articles, Hamermesh (2013) documented this trend, demonstrating that empirical work has become considerably more prevalent in leading economics journals since the 1960s. Angrist et al. (2017) expanded on this research by utilising machine learning techniques to examine 135,000 journal articles published in 80 works frequently cited in the American Economic Review. The authors contend that the shift toward empirical research is not due to more empirical areas replacing more theoretical ones, but rather that every subject is becoming more focused on empirical methods. From this perspective, the emergence of big data offers immense potential, owing to the vast quantities, variety, and the ability to link various datasets.

This may lead to more precise modelling of social and economic issues and enhanced causal inference, empowering researchers to address pertinent questions.

The modelling of the UK housing market has been ongoing since the 1970s (Ball 1973; McAvinchey and Maclennan 1982), with a substantial portion of the data either aggregated to broader geographic regions or districts, or alternatively, linked to individual properties in specific cities. Aggregate mortgage data samples, primarily from building societies, have been extensively employed (Alexander and Barrow 1994; Cook 2003; Hudson et al. 2018). However, while these datasets lack local granularity, they also introduce potential biases due to their inherent limitations and small sample sizes. Conversely, local estate agent survey data could provide more comprehensive micro-level housing information, enabling detailed local housing analyses (Orford 2010). Nevertheless, such datasets remain scarce.

The current research on house price variation in the UK is impeded by the absence of an open and comprehensive housing database that includes transaction prices, listing prices, and property attributes, such as floor size, number of habitable rooms, location, and so forth. The limited data availability, coupled with disparities in price derivation methods, reporting formats, and geographical coverage (Ciarlone 2015), presents numerous challenges in conducting empirical analyses of residential property prices. Although some studies have explored the connection between open source data to alleviate these constraints (Chi et al. 2021; Jonathan et al. 2018; Powell-Smith 2017), they only made use of at most two administrative open datasets. Modelling based on these linked data might still encounter endogeneity issues due to inadequate information from data sources, complicating the identification of causal effects and hindering the understanding of the market (Hayunga and Pace 2019; Huang and Milcheva 2020). These limitations highlight the necessity for standardised and comprehensive data to facilitate robust analysis and comprehension of housing markets.

To address this gap, I propose a framework for data collection and integration, subsequently creating several micro-level big datasets on residential property transactions in England. This approach enabled me to gain new insights into two key aspects of understanding the market: i) resolving a long-standing puzzle concerning the relationship between price and time on the market; and ii) examining the effects of transaction tax reduction on trading volumes, prices, and liquidity. To the best of my knowledge, these datasets are among the most comprehensive employed in similar research, offering innovative solutions to these questions and providing a unique opportunity for a thorough understanding of the market. The datasets cover residential property transactions in England recorded by the Land Registry between January 2018 and March 2021, ranging from one-third to 92% of all transactions at full market price. The datasets included information on property listings and transactions, physical and energy performance attributes, and council tax information.

In order to compile this multi-source information and link relevant details, a series of computer programmes and algorithms was developed using the R programming language.[1] While machine learning methods, often associated with big data, were not utilised in this study, other econometric methods were chosen for their alignment with the causal inference framework required to address my research questions. The emphasis was placed on employing appropriate methods to establish causality and offer insights based on real-world data. The discussion of these choices is presented in the concluding section of this chapter.

This research heavily relied on the utilisation of open data, capitalising on the UK government's strong commitment to the open data movement. Over the past decade, the government has actively published high-quality datasets and promoted the use of open data among businesses and researchers. In 2012, the government implemented the 'Open by Default' policy introduced in their "Open Data White Paper"[2], which aimed to increase transparency and accountability while generating positive economic outcomes. With the advent of big data, the government has also provided various support mechanisms for academic researchers to embrace the trend. In June 2014, the government introduced new copyright exceptions[3]. Consequently, academic re-

[1] R is a free software environment for statistical computing and graphics: https://www.r-project.org/.

[2] 'Open Data: unleashing the potential': https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

[3] In summary, scraping copyright-protected material from the web is allowed if one has access to

searchers have been empowered to use copyright-protected material for their research purposes without the need for permission from the copyright owner. This has opened up new opportunities for researchers in various fields such as economics, sociology, and geography to gather data from diverse sources, including traditional websites and social media platforms, through web scraping, text mining, and data mining techniques. The remaining challenges for researchers utilising big data primarily involve technical barriers, such as website rate limits and the issue of assembling data from various sources without a universal identifier.

1.2 Contributions

In Chapter 2, I present a novel framework for data collection and integration, which involved creating several micro-level big datasets from four main sources. Two of these sources, the Price Paid Data[4] (PPD) and the Energy Performance Certificates data[5] (EPC), are open administrative datasets available to the UK public. The Land Registry's PPD is a comprehensive dataset containing address-level records of all property sales in England and Wales that are sold for value and registered with them from 1995 to the present. The EPC data is hosted by the Department for Levelling Up, Housing and Communities (DLUHC). An EPC must be issued or renewed by law when a property is built, sold, or rented, and a surveyor assesses the property and reports information about property characteristics in addition to energy performance.[6] Both datasets are well-organised in table form and can be downloaded directly from the government's website.

In addition to the aforementioned datasets, I utilise publicly accessible council tax information from the Valuation Office Agency (VOA).[7] This data has not been previously employed in related research. Particularly important in this study is the

it and if the research is non-commercial. Details can be found from the British Intellectual Property Office: https://www.gov.uk/guidance/exceptions-to-copyright.

[4] Contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0.

[5] The data can be accessed here: https://epc.opendatacommunities.org

[6] See, https://epc.opendatacommunities.org/docs/guidance

[7] Council tax band lookup portal: https://www.gov.uk/council-tax-bands

Council Tax Band (CTB), which plays a crucial role in identifying causal effects. This is because all properties are valued and placed into bands on the same basis, and the assessment is exogenous to all participants. Consequently, the CTB effectively resolves several endogeneity issues in modelling various research questions. However, there is no direct method to download the data, so I develop a web scraping process to collect the CTB data.

In addition to the government data sources, I also obtain monthly new listings of residential properties from Zoopla's API[8]. This data includes several vital pieces of information, such as the initial listing date and price, the final listing price, and whether the listing is chain-free. To collect the listings data through the API, I develop a programme and gathered this data on a monthly basis since 2018.

However, none of the datasets mentioned above have a universal identifier. Therefore, I develop a series of text matching and unique identifier matching algorithms to generate multiple big datasets for the studies presented in Chapters 3 and 4.

In Chapter 3, I revisit the long-standing puzzle regarding the relationship between price and time on the market (TOM) in housing studies. Although housing plays a crucial role in household portfolios, its heterogeneity and illiquidity render the accurate valuation of residential transactions a complex task for market participants. The Hedonic model (Rosen 1974) proposes that properties are valued based on their utility-bearing attributes, such as physical features and location-specific amenities and services. Nevertheless, even after accounting for these attributes, prices remain dispersed instead of uniform within the local market (He et al. 2017).

The search and matching theory of housing markets establishes a framework to comprehend the process of discovering a property's true value and the resulting equilibrium market price. According to this theory (Anglin et al. 2003; Krainer and LeRoy 2002; Wheaton 1990), the price and TOM are simultaneously dependent on the probability of sale, implying a positive correlation between the two (Hayunga and Pace 2019). However, the empirical evidence supporting this relationship is am-

[8] The Zoopla API: `https://developer.zoopla.co.uk/home`. Zoopla is the second largest property listing website in the UK and claims 70% coverage of UK residential listings since 2008. The API is no longer freely accessible since January 2023

biguous. Estimating this relationship is influenced by various factors, including the overall state of the housing market, the location and characteristics of the property, and the specific circumstances of the seller. The impact of TOM on price and vice versa has been the focus of extensive research over the past three decades. Before 2015, a significant number of studies presented inconclusive findings on this subject (Benefield et al. 2014; Johnson et al. 2007). Although several recent investigations have directly addressed this puzzle, a consensus has yet to be reached (Dubé and Legros 2016; Hayunga and Pace 2019; He et al. 2017).

Expanding on prior literature, I identify two primary causes of the discrepancies in empirical outcomes concerning the relationship between price and TOM. The first cause is a model identification issue arising from endogeneity due to the joint determination of price and TOM. The second cause is the absence of a variable accounting for overpricing, often excluded in earlier studies because of data constraints. By capitalising on my extensive data, I tackle the first issue by identifying two innovative IVs through a 2SLS estimation approach, and the second issue by incorporating a newly devised measure of overpricing in the model. My findings corroborate a positive association between price and TOM using a simultaneous equation model, aligning with search theory.

Unexpectedly, the 2SLS findings indicate that, all else being equal, properties listed as ”chain-free” sold for 4-5% less on average compared to ”in-chain” properties, even though ”chain-free” listings are frequently perceived as a selling advantage in practice, providing a more flexible and efficient buying process. Upon investigating the mediation effect of the initial listing price, I discover that chain-free sellers, not required to sell their current property before acquiring a new one, tend to set lower initial asking prices and accept reduced transaction prices.

Levitt and Syverson (2008) demonstrated that agents tend to establish a lower initial listing price to expedite the sale. In contrast, ”in-chain” sellers are generally more financially constrained, and the pace of their sales relies on the progress of other sales within the chain. Consequently, agents are less inclined to convince these sellers to set a lower initial list price due to their heightened risk aversion, and rapid sales are unattainable owing to the chain's inherent nature. As a result, "in-chain" sellers usually have a higher initial listing price and encounter fewer principal-agent issues compared to "chain-free" sellers. Thus, the agency costs problem is a key factor contributing to a lower initial listing price for properties not involved in a chain.

In Chapter 4, I investigate the effects of property transaction tax changes, specifically the 2020 stamp duty land tax holiday (SDLT), on the housing market in the UK. Property transaction taxes are levied on purchasing real estate in many countries. This tax is typically based on a percentage of the sale price of the property, and it is generally paid by the buyers at the time of the sale. In the UK, the stamp duty rate varies based not only on the value of the property but also the type of buyer, with different rates applying to first-time buyers, second-home buyers, investors, and other criteria.

Stamp duty in the UK serves dual purposes of generating government revenue and regulating the housing market. The funds raised from stamp duty finance public programs and services like education, healthcare, and infrastructure development. It also helps prevent housing bubbles by making it difficult for speculators to excessively buy and sell, promoting long-term ownership and fostering stable communities. Property owners are encouraged to contribute to the public good rather than solely for personal gain.

However, stamp duty is unattractive for reducing the expected benefits of buyers and sellers by discouraging trades, making it harder for properties to be held by those who value them most. Previous researches have criticised transaction taxes for their negative impact on mobility, hindering people from relocating for better opportunities and resulting in negative effects on employment, productivity, etc. (Hilber and Lyytikäinen 2017; Van Ommeren and Van Leuvensteijn 2005). In addition, the frequency of property transactions varies greatly across regions and households, but there's no strong justification for imposing excessive taxes on frequently traded residential properties (Adam 2011).

In December 2014, the UK government reformed the SDLT from a "slab" system to a "slice" system, resulting in a stamp duty reduction for most taxpayers[9]. Despite numerous studies on the "slab" SDLT, little is known about the effects of the new "slice" system. This study aims to address this gap by thoroughly analysing the impact of changes in the new progressive tiered tax system on the residential market, using the 2020 SDH as a quasi-natural experiment.

In June 2020, as part of job creation measures, a temporary reduction in stamp duty was introduced with immediate effect in response to the stagnant housing market during the early stages of the COVID-19 pandemic. This policy intervention provides a quasi-natural experimental setting, which I leverage in this study to evaluate the effects of the reduction of "slice" stamp duty on prices, trading patterns, and liquidity in the housing market.

Theoretically, I propose a Nash bargaining model to explain the "slice" SDLT, and show that the tax holiday can cause an increase in prices and sellers will have more surplus if they trade during SDH. Empirically, I estimate a series of DiD models and find that, on average, the SDH caused a 53% increase in housing transactions, a 60% rise in listings and an over 2% increase in transaction prices. Additionally, I observe that sellers had stronger bargaining power as the SDLT holiday deadline approached. The entire tax savings from the SDH was passed on to sellers in the form of increased prices, reducing affordability for first-time buyers and home movers replacing their main residence. The results also provide evidence that market participants used the SDH to relocate away from the highly urbanised polycentric areas during the Covid-19 pandemic. My findings show that while a SDH can stimulate market activity during an economic downturn and enable the housing market to adjust to changing conditions quickly, it also has an unintended consequence of reducing housing affordability.

[9] Preliminary Assessment of 2014 Residential SDLT 'Slice' Reforms: https://www.gov.uk/government/publications/preliminary-assessment-of-2014-residential-sdlt-slice-reforms

1.3 Incorporate Machine Learning for Future Research

In the realm of big data, machine learning (ML) emerges as a potent tool for extracting insights and making predictions from complex datasets, with applications spanning image recognition, natural language processing, and predictive modelling. While vast data volumes open doors for exploration with ML, this thesis prioritises econometric methods for the specific research questions explored. This choice recognises the historical strengths of econometrics in causal inference, especially when considering the development of the research questions in this thesis. As mentioned in the previous section, chapter 2 aims to solve issues presented in econometric methods in previous research, and the observational data in chapter 3 naturally forms a quasi-natural experiment that is a good fit for the DiD method.

Traditionally, ML algorithms, such as random forests, lasso, ridge, deep neural nets, boosted trees, and various hybrids and ensembles of these methods, are designed to utilise the correlations between variables and patterns in data to make predictions[10]. Most model selections prioritise high predictive power through the cross-validation method, often without inherently considering causality (Athey and G. W. Imbens 2019). Furthermore, ML algorithms often entail complex mathematical models or black-box algorithms, such as those related to neural networks, making it challenging to interpret and comprehend the underlying mechanisms being studied. This poses an obstacle to using machine learning for causal inference since a deep understanding of the underlying mechanisms is often required to identify and interpret the causal relationships between variables.

However, there have been some advances in ML for causal inference in recent years. Athey and G. Imbens (2016) propose a modified tree model from ML to do valid inference for the causal effects in randomised experiments and in observational studies satisfying unconfoundedness. Griffin et al. (2017) demonstrate that ML meth-

[10] This has been discussed in many textbooks that cover ML methods alongside more traditional statistical methods, e.g., Hastie et al. (2009) and James et al. (2013).

ods (boosted regression) can lead to good estimates of the propensity score for the matching method in causal inference. These highlight the idea that future research can combine ML approaches for the prediction component of models with causal approaches. For example, in this thesis, ML methods can be adopted for further exploration. This could involve improving the estimation of the overpricing proxy and creating a novel instrumental variable for 2SLS estimation. This can be achieved through the utilisation of ML methods and big data, including the prediction of property prices.

Chapter 2

The Construction of the Big Data

Since the 1970s, researchers have been modelling the UK's housing market (Ball 1973; McAvinchey and Maclennan 1982) by utilising available information, either combined into larger geographic regions or associated with specific properties in certain areas. For example, building society mortgage data, including aggregate sample mortgage data, has been heavily employed in research (Alexander and Barrow 1994; Cook 2003; Hudson et al. 2018). However, these datasets have limitations, including a lack of detailed information and the potential for biases due to their limited sample size. While local estate agent survey data offers the potential for detailed micro-level insights into the housing market (Orford 2010), such datasets are not readily available. The availability of the Land Registry PPD as open data since 2013 has brought about a transformative effect on research into the UK housing market (Cooper et al. 2013; Gray 2012). However, one of its significant shortcomings is the absence of physical property characteristics such as floor size (Orford 2010).

The current research on residential house price variation in the UK is hindered by the lack of an open and comprehensive house price database that contains transaction prices and property attributes (Chi et al. 2021). The limited availability of data, combined with variations in the method of price derivation, reporting formats, and geographical coverage (Ciarlone 2015), poses significant challenges for conducting empirical analysis of residential property prices. Furthermore, insufficient data can lead to endogeneity issues that impede the identification of causal effects, thereby hindering market understanding (Hayunga and Pace 2019; Huang and Milcheva 2020). These limitations underscore the critical need for standardised and comprehensive data collection methods and frameworks to enable robust analysis and understanding of housing markets.

The main objective of this chapter is to bridge the gap in constructing large datasets on the residential market by utilising several openly accessible data sources. This study leverages the UK government's commitment to the open data movement and research-related copyright exceptions by combining information from four primary data sources, namely PPD, EPC data, Zoopla's listings data, and CTB data, as summarised in Table 2.1. To construct these large datasets, a series of programs and algorithms have been developed, which are listed in Table 2.3 at the end of this chapter.

The following sections provide an introduction to each data source's specifics, significance, and availability, along with an overview of the process of linking the information from these sources to construct several property-level large datasets. The smallest dataset constructed contains information on one-third of the population's residential transactions, while the largest dataset encompasses over 92% of all transactions recorded by the Land Registry in the sample period. To the best of my knowledge, these datasets are among the most comprehensive used in similar studies and offer unique insights into the UK residential market. These insights are further explored in Chapters 3 and 4. The descriptions of variables used in these two chapters are listed in Table 2.2 at the end of this chapter.

2.1 Data Sources

Price Paid Data

Constructing comprehensive residential datasets requires the integration of multiple data sources. A crucial component is the price paid transaction data, which provides detailed information on all property sales. This data is collected by the Land Registry,

Table 2.1: The Main Data Sources

Data Source	Description	Availability
Price Paid Data (PPD)	It contains information on the transactions of residential properties in England and Wales. It includes the date of the sale, the property address, the type of property, the sale price, etc.	This data is collected by the Land Registry and is made available to the public.
Energy Performance Certificates Data (EPC)	EPC is required whenever a property is built, sold, or rented. The data contains information on the property's energy efficiency and physical characteristics, such as the total floor area, number of habitable rooms, and type of glazing.	This data is collected by the Department for Levelling Up, Housing & Communities and is made available to the public.
Zoopla's Listings Data	Zoopla is a well-known online property website in the UK offers a wide range of information on properties available for sale and rent, including descriptions, photos, and pricing details.	It can be retrieved from Zoopla's open API using a specially developed computer program.
Council Tax Band Data (CTB)	It contains the property address and utilises a scale ranging from A to H (in England and Scotland) or A to I (in Wales) to categorise properties based on their value.	The CTB for a property can be searched on the government's website using its postcode.

a government agency responsible for maintaining a record of property ownership and transactions. The PPD is openly available to the public and can be directly downloaded from a government website[1]. The PPD includes records for all property sales in England and Wales that have been sold for value and registered with the Land Registry from 1995 to the present. It is a valuable resource for understanding trends and patterns in the housing market and for conducting research on the economic and social factors that influence the value of residential properties.

The PPD typically provides the following information on the sale price of a property: the postcode and full address of the property, the type of property (e.g., detached, semi-detached, terraced, flats/maisonettes, or other), the date of the transfer (i.e., the date on which the sale was completed, as stated on the transfer deed), the tenure of the property (e.g., freehold or leasehold[2]), and a dummy variable for newly-built properties. Additionally, the Land Registry assigns a unique transaction identifier to each record in the PPD. However, this identifier is only unique within the PPD and cannot be used to link the PPD with other datasets. Therefore, it is necessary to use alternative methods, such as the property address or additional information about the property, to establish a link between the PPD and other datasets.

It is worth noting that the PPD includes sales that were not for the full market value, such as transfers under a power of sale or repossessions, buy-to-lets (identified by a mortgage for landlords who want to buy property to rent it out), and transfers to non-private individuals. These types of sales are recorded in the PPD as additional price paid transactions and can be identified by filtering the data to include only those records with a PPD Category Type of ”B”. For the purpose of this research, it

[1] Download the Price Paid Data: `https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads`

[2] In the UK, there are two types of property tenure. Not all properties are freehold, although ownership is transferred to the buyer. Apartment units and some houses are leaseholds. This is due to historical reasons dating back to when the land was not privately owned, and leaseholders would pay ground rent to freeholders. Ground rent is still paid for leasehold properties, although the amount can vary widely across units. A leasehold property typically has a lease of 100 years, but it can range from a few years up to 999 years. The lease can be extended at the request of the owner for a fee, which makes the ownership of a leasehold property comparable to a freehold. In general, a leasehold property is considered less preferable than a freehold property and may sell for less, all else being equal(Lai and Milcheva 2021).

is necessary to exclude these additional price paid transactions and only include sales with a full market value in the analysis. This can be achieved by filtering the data to include only those records with a PPD Category Type of ”A”, which indicates a standard price paid entry for a single residential property sold for value.

Energy Performance Certificates Data

The EPC data is another important source of information on residential properties. Legally mandated in the UK, EPCs are documents that provide insights into the energy efficiency of a given property, which must be obtained upon construction, sale, or rental. This governmental measure was introduced in 2008 with the aim of curtailing energy consumption and reducing greenhouse gas emissions. Each certificate is assessed by qualified and accredited energy assessors who visit the property and gather information about key elements such as cavity wall insulation, floor and loft insulation, boilers, radiators, heating controls, windows, and other relevant details based on a standard assessment procedure recommended by the government.

The EPC data contains information about the energy efficiency of a property, including ratings on a scale from A to G, with A being the most energy efficient and G being the least efficient. Each certificate is valid for ten years and can provide high-quality information about a property. EPC data can be used to understand the environmental impact of different properties and to identify opportunities for energy efficiency improvements.

Beyond providing insight into the energy efficiency of a given property, EPC data includes several other essential physical attributes of properties[3], such as the total floor area, the number of habitable rooms[4], the number of open fireplaces, the

[3] Guidance of EPC data: `https://epc.opendatacommunities.org/docs/guidance`

[4] The definition of habitable rooms includes various living spaces such as the living room, sitting room, dining room, bedroom, study, and similar spaces, along with a non-separated conservatory. Additionally, a kitchen/diner with a separate seating area having space for a table and four chairs is also considered a habitable room. If a non-separated conservatory has an internal quality door between it and the dwelling, it adds to the habitable room count. However, excluded from the room count are rooms used solely as a kitchen, utility room, bathroom, cloakroom, en-suite accommodation, and other similar spaces. Additionally, any hallway, stairs, landing, or room without a window is not included in the habitable room count.

number of extensions, built form classification (e.g., detached, semi-detached, mid-terrace, and end-terrace), and property type classification (e.g., house, bungalow, flat, and maisonette). These details provide an accurate and detailed view of properties compared to similar information provided in listings by real estate agents and are useful as high-quality control variables in models.

In addition to these physical attributes, the EPC data also includes detailed address information for each property, which facilitates data matching with other sources. As of November 2021, the EPC data collected by the DLUHC has also included the Unique Property Reference Number (UPRN) for each property[5]. The UPRN is a one-of-a-kind identifier that allows for cross-referencing between different datasets and simplifies data linking. With the inclusion of UPRN to the EPC data, researchers can now more effortlessly merge and analyse multiple data sources, allowing for more comprehensive analyses of the housing market.

This research incorporates various variables from EPC data, such as the total floor area, number of habitable rooms, number of open fireplaces, number of extensions, current energy efficiency, potential energy efficiency, current and potential environmental impact, and current and potential energy consumption. Additionally, the EPC property type classifications are preferred over type information in PPD when possible, as EPC data provides more detailed information. Properties are classified into four types, including house, bungalow, flat, and maisonette, combined with four built forms: detached, semi-detached, mid-terrace, and end-terrace. The EPC variables offer useful insights into the property's quality and reduce information asymmetry between buyers and sellers, as stated by Parkinson et al. (2013) and Aydin et al. (2019). In the commercial real estate sector, studies have shown that buildings with better energy efficiency fetch higher gross rents, enabling landlords to reap a premium, as reported by Szumilo and Fuerst (2015).

[5] More details: https://news.opendatacommunities.org/energy-performance-certificates-now-include-uprn/

Listings Data from Zoopla

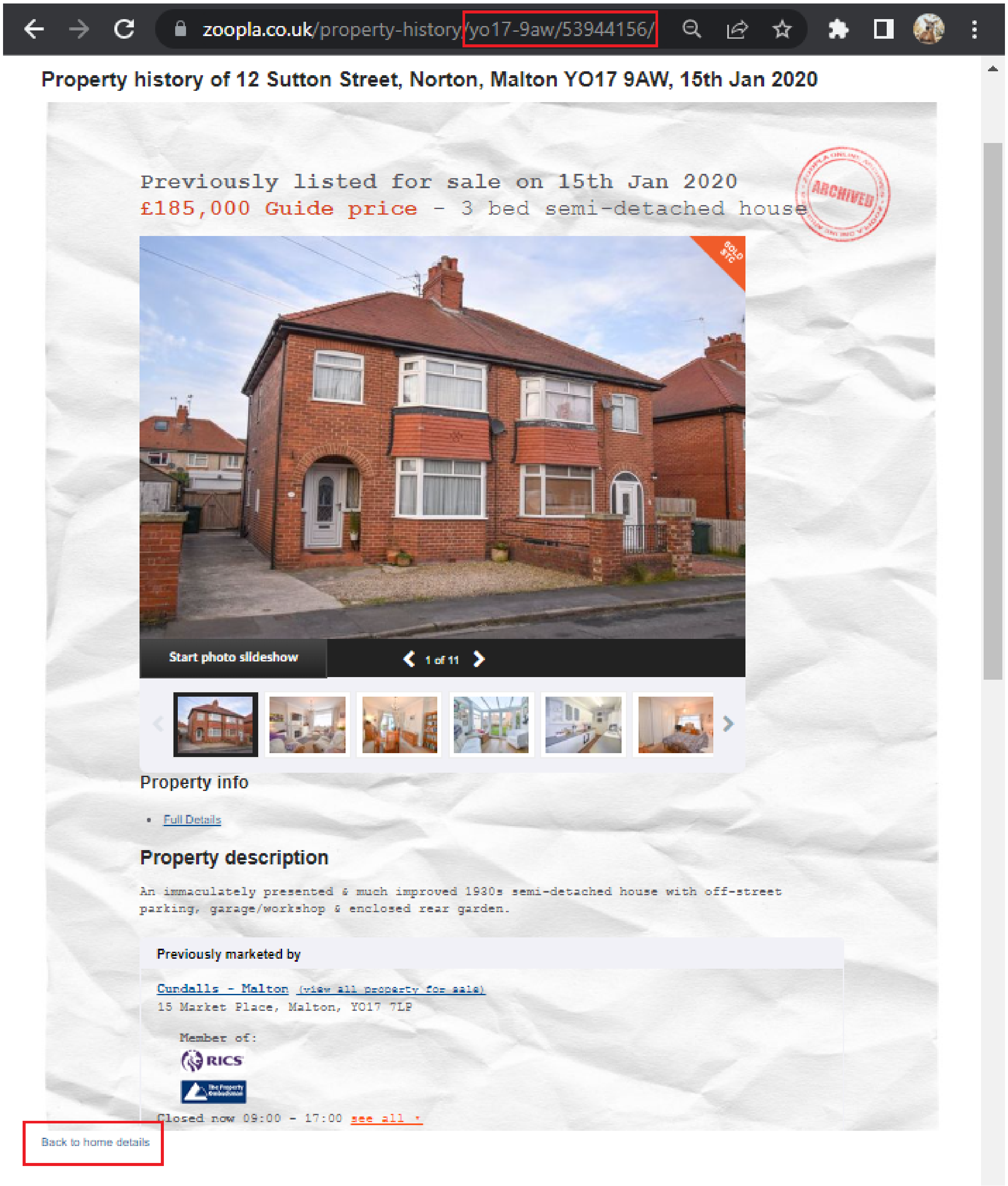
In addition to the PPD and EPC data, this research incorporates the listings data from Zoopla, one of the largest online property websites in the UK, which provides a wide range of information on properties available for sale and rent, including descriptions, photos, and pricing details. This data can offer valuable insights into the characteristics and availability of various types of properties on the market, as well as trends and patterns in the housing market, such as changes in demand for specific types of properties or shifts in pricing patterns.

Specifically, the listings data tracks the initial asking price and any subsequent changes in the asking price, along with the corresponding dates of those changes. Additionally, the data allows for the identification of the number of bathrooms, bedrooms, floors, and reception rooms. Another advantage of using this data is the ability to extract valuable information from the textual descriptions of properties using text mining techniques. This provides researchers with additional details about properties that are absent in the other two data sources. For example, text mining techniques were applied to create variables that reveal whether a property includes a garden, garage, or driveway, as well as an indicator if a property is listed as ”chain-free” (more details will be discussed in Section 2.3).

Zoopla has improved accessibility to its data by implementing an open Application Programming Interface[6] (API). In order to support this research, a program was created to systematically and automatically collect listings data using the API. This approach ensures efficient analysis of a large volume of data, enabling examination of changes and trends in the housing market over time with a reliable and up-to-date data source.

However, the listings data obtained from the API does not provide detailed addresses for each record, which limits its potential for matching with data from other sources. To overcome this limitation, a two-step web-scraping procedure was developed. In the first step, a program was written to use an open-sourced geo-location

[6] An API provides external developers with a set of programming instructions that allow them to access the data and functionality of a website or service in a controlled manner.



Property history of 12 Sutton Street, Norton, Malton YO17 9AW, 15th Jan 2020

Previously listed for sale on 15th Jan 2020
£185,000 Guide price - 3 bed semi-detached house

Property info

- Full Details

Property description

An immaculately presented & much improved 1930s semi-detached house with off-street parking, garage/workshop & enclosed rear garden.

Previously marketed by

Cundalls - Malton (view all property for sale)
15 Market Place, Malton, YO17 7LP

Member of:

Closed now 09:00 - 17:00 see all

Figure 2.1: Retrieve detailed address with listing ID and postcode
Note: The program first constructs the web page address by concatenating a string with the postcode and listing ID. It then visits the web page and collects the address information displayed at the top of the page.

API[7] to retrieve the full postcode for each property based on the latitude and longi-

[7] Postcodes.io is an open-sourced project that allows developers to search, reverse geocode and extract UK postcode and associated data: `https://postcodes.io/`

tude coordinates provided in the listings data. In the second step, as shown in Figure 2.1, a web-scraping program was created to gather the address information of each property using its corresponding Zoopla listing ID (which existed in the listings data) and the corresponding postcode retrieved from the previous step.

In 2021, Zoopla began incorporating the UPRN into its listings data, which simplifies linking the data with other datasets that also use the UPRN, such as the EPC data. By utilising the UPRN as a standard identifier, multiple data sources can be merged and analysed more thoroughly and accurately. However, the Zoopla API does not contain the UPRN in its listings data. Instead, it can be found on the property details page for each listing, which can be accessed via a new URL provided by Zoopla[8]. Therefore, to combine the Zoopla listings data with other datasets that utilize the UPRN, a web-scraping program was developed, as illustrated in Figure 2.2, to collect the UPRN for each record in the listings data.

It is crucial to emphasise that the address-related information and the UPRN acquired through web-scraping programs are exclusively utilised for matching purposes in our research, adhering strictly to the ”Copyright, Designs and Patents Act 1988”[9].

Council Tax Band Data

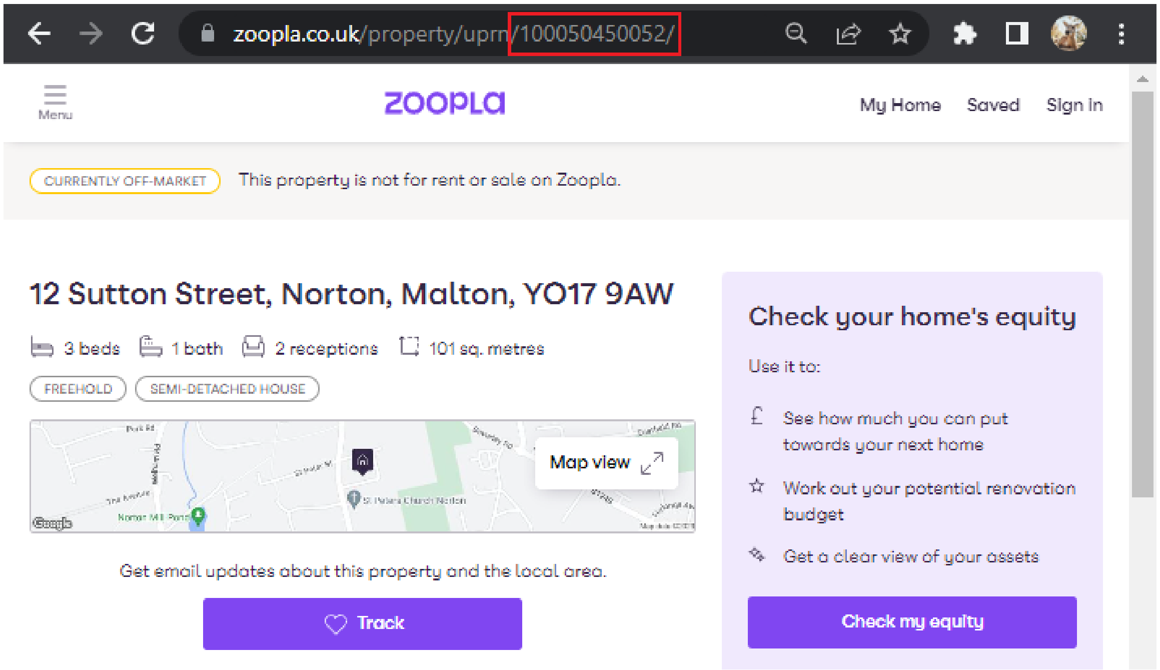
Ultimately, council tax data is integrated into extensive residential datasets. This information, gathered by local authorities, comprises details on the address and council tax band for each property.

The CTB is a classification system employed to ascertain a property's value based on various criteria, including size, layout, character, location, and alterations in usage. This valuation is founded on the probable selling price of the property on the open market as of 1 April 1991 in England[10]. The Valuation Office Agency, a governmental body, is tasked with determining CTB values, which are categorised using a scale that

[8] https://help.zoopla.co.uk/hc/en-gb/articles/4409413698321-Why-does-my-address-look-different

[9] Exceptions to copyright of the owner according to the 'Copyright, Designs and Patents Act 1988' in the UK. https://www.gov.uk/guidance/exceptions-to-copyright

[10] https://www.gov.uk/guidance/understand-how-council-tax-bands-are-assessed



zoopla.co.uk/property/uprn/100050450052/

CURRENTLY OFF-MARKET This property is not for rent or sale on Zoopla.

12 Sutton Street, Norton, Malton, YO17 9AW

3 beds 1 bath 2 receptions 101 sq. metres

FREEHOLD SEMI-DETACHED HOUSE

Get email updates about this property and the local area.

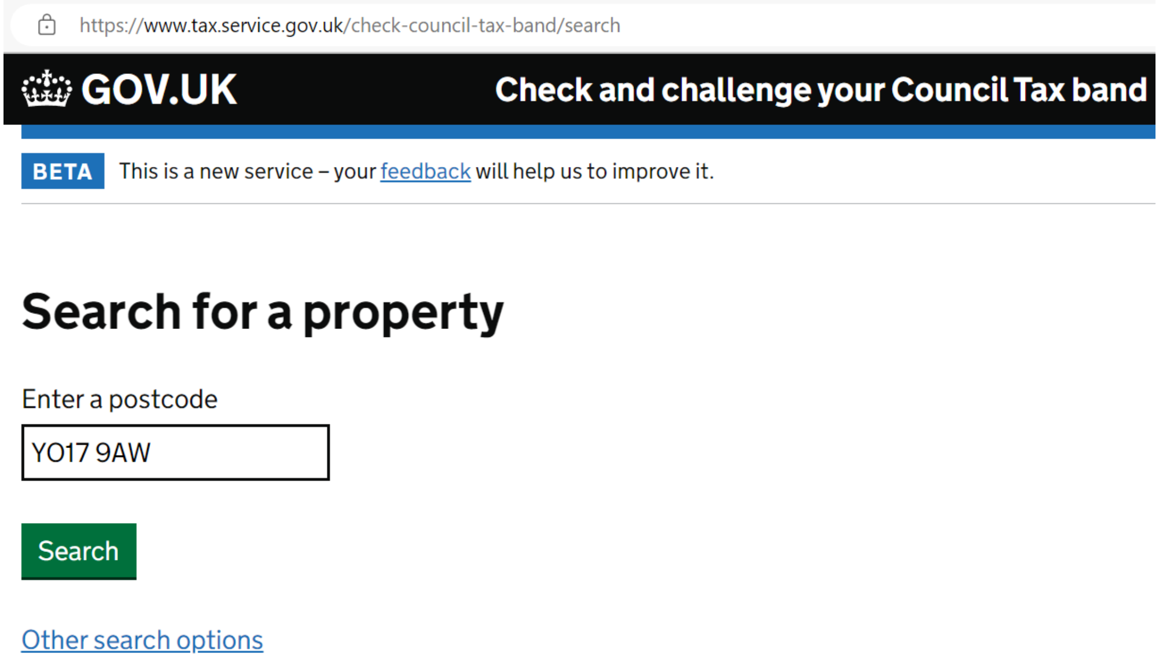
Check your home's equity

Use it to:

- See how much you can put towards your next home
- Work out your potential renovation budget
- Get a clear view of your assets

Figure 2.2: Retrieve UPRN from Zoopla web page

Note: The program emulates a click on the ”Back to Home Details” link, located in the bottom left corner of Figure 2.1 and marked within a red box. Following this, it extracts the UPRN from the redirected page, as highlighted by the red box in the same figure.



https://www.tax.service.gov.uk/check-council-tax-band/search

Check and challenge your Council Tax band

This is a new service – your feedback will help us to improve it.

Search for a property

Enter a postcode

YO17 9AW

Figure 2.3: Council tax band lookup portal

spans from A to H (in England and Scotland) or A to I (in Wales), where A signifies the lowest value and H/I denotes the highest value. Local councils annually establish council tax rates for each valuation band, with properties in higher bands generally subject to increased council tax rates. It is worth noting that property owners can contest their CTB if they believe their home has been incorrectly categorised within a specific valuation band.

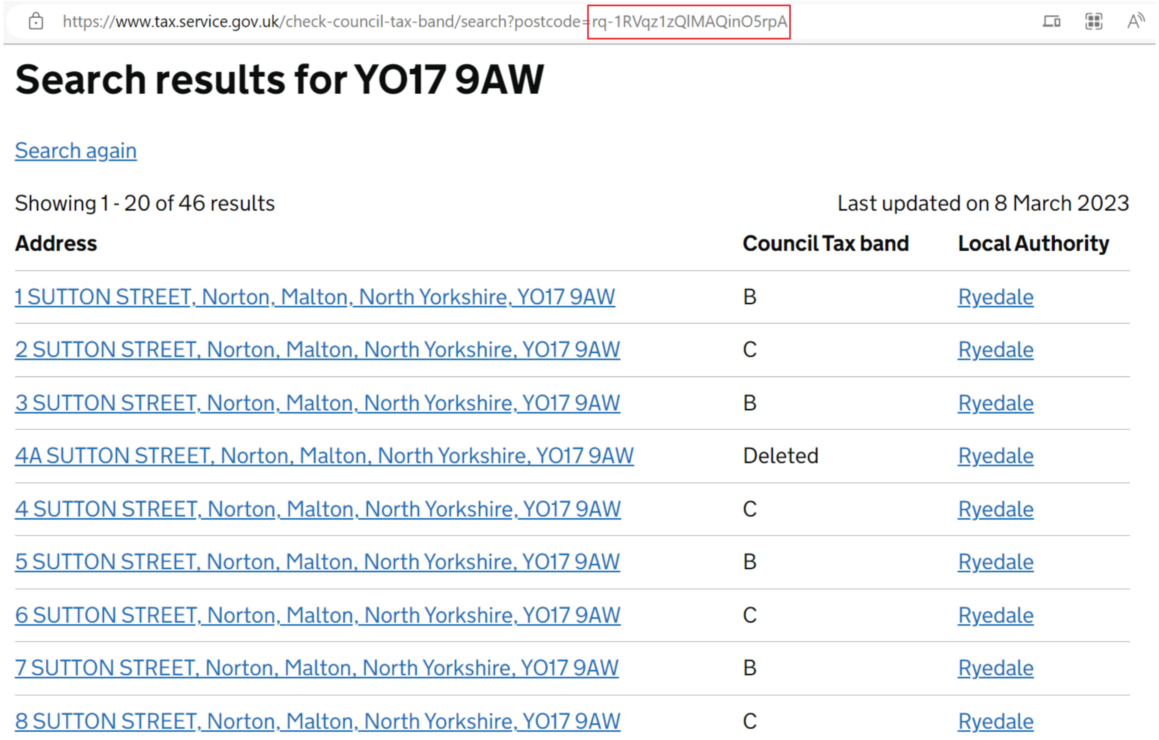
Council tax data serves as a valuable resource for researchers, primarily because the CTB is determined by an independent third party, the VOA. Consequently, it remains exogenous to all property transaction participants, including buyers, sellers, and agents. This ensures that the CTB is not subject to influence from the actions or decisions of these individuals, rendering it a dependable indicator of a property's value. As a result, the CTB is a vital variable in numerous models, offering an impartial and unbiased measure of a property's value that can be employed to analyse and compare properties across various regions in the market.

The Council Tax Band (CTB) for each property is publicly accessible on the government website[11] (Figure 2.3). To integrate the CTB into this research, I developed a web-scraping program to obtain the CTB for each property from the CTB lookup portal. As illustrated in Figure 2.4, the web address for the CTB results of a postcode area comprises a unique fixed string and an encoded postcode string. Since the encoding method is not publicly disclosed, the web address of the search results page for each postcode area is unpredictable, making direct data scraping infeasible.

Consequently, the program employs the Selenium project[12] to simulate manual web browsing. This includes inputting the postcode into the search box (as displayed in Figure 2.3) and activating the search button. Subsequently, it reads and compiles the CTB data presented on the results page.

[11] Council tax band lookup portal: `https://www.gov.uk/council-tax-bands`

[12] Selenium is a widely-used open-source framework for automated testing of web applications. It provides a suite of tools for automating web browsers and performing various testing tasks such as clicking buttons, filling out forms, and navigating between pages.



https://www.tax.service.gov.uk/check-council-tax-band/search?postcode=rq-1RVqz1zQIMAQinO5rpA

Search results for YO17 9AW

Search again

Showing 1 - 20 of 46 results

Last updated on 8 March 2023

Address	Council Tax band	Local Authority
1 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	B	Ryedale
2 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	C	Ryedale
3 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	B	Ryedale
4A SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	Deleted	Ryedale
4 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	C	Ryedale
5 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	B	Ryedale
6 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	C	Ryedale
7 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	B	Ryedale
8 SUTTON STREET, Norton, Malton, North Yorkshire, YO17 9AW	C	Ryedale

Figure 2.4: Council tax band searching

Note: The web address for the CTB results in a postcode area is composed of a unique fixed string and an encoded postcode string, highlighted within the red box at the top of the figure. To gather the data, a program was devised to emulate manual web browsing, encompassing the input of the postcode in the search box and the activation of the search button (as illustrated in Figure 2.3). Subsequently, it reads and compiles the council tax data presented on the results page.

2.2 Data Integration

To generate comprehensive housing datasets, information from the four previously mentioned sources was merged using both address matching and unique identifier matching algorithms, specifically designed for this research. The data integration process is succinctly depicted in Figure 2.5. Different combinations of the four sources resulted in three new comprehensive datasets: PPD-EPC-CTB, Listings-EPC-CTB, and Listings-PPD-EPC-CTB. These datasets encompass information on property transactions that took place between January 1, 2018, and March 31, 2021, and serve as the basis for analysis in the subsequent two chapters.

Price Paid Data - Energy Performance Certificates Data

In the first stage of the data integration process, a specialised algorithm (Match Algorithm 1 in Figure 2.5) was designed to link the records in the EPC data with the property transactions recorded in the PPD. This algorithm utilised detailed address information and postcodes as common identifiers.

This algorithm first converted the address information of each record into a string, replacing most words with their commonly used abbreviations (e.g. ”Street” or ”St.”) and stemming[13] the words to avoid issues with different spellings or variations of the same address. The numerical component was also extracted to improve the probability of correct matching.

The algorithm then compared the records in the PPD and EPC datasets based on their postcodes and calculated a measure of string similarity[14] between the address strings in the pre-matched results, which was represented as a number between 0 and 1. A value of 0 indicated no shared characters, while a value of 1 indicated a perfect match. The algorithm also calculated an indicator for whether the address-associated numbers were identical in the pre-matched results. If two records from the PPD and

[13] Stemming is a technique used to extract the base form of the words by removing affixes from them, similar to cutting down the branches of a tree to its stems. For example, the stem of the words eating, eats, eaten is eat.

[14] It uses the Jaro-Winkler string similarity algorithm, which is also used by the UK's National Health Service data science team for similar address matching tasks.

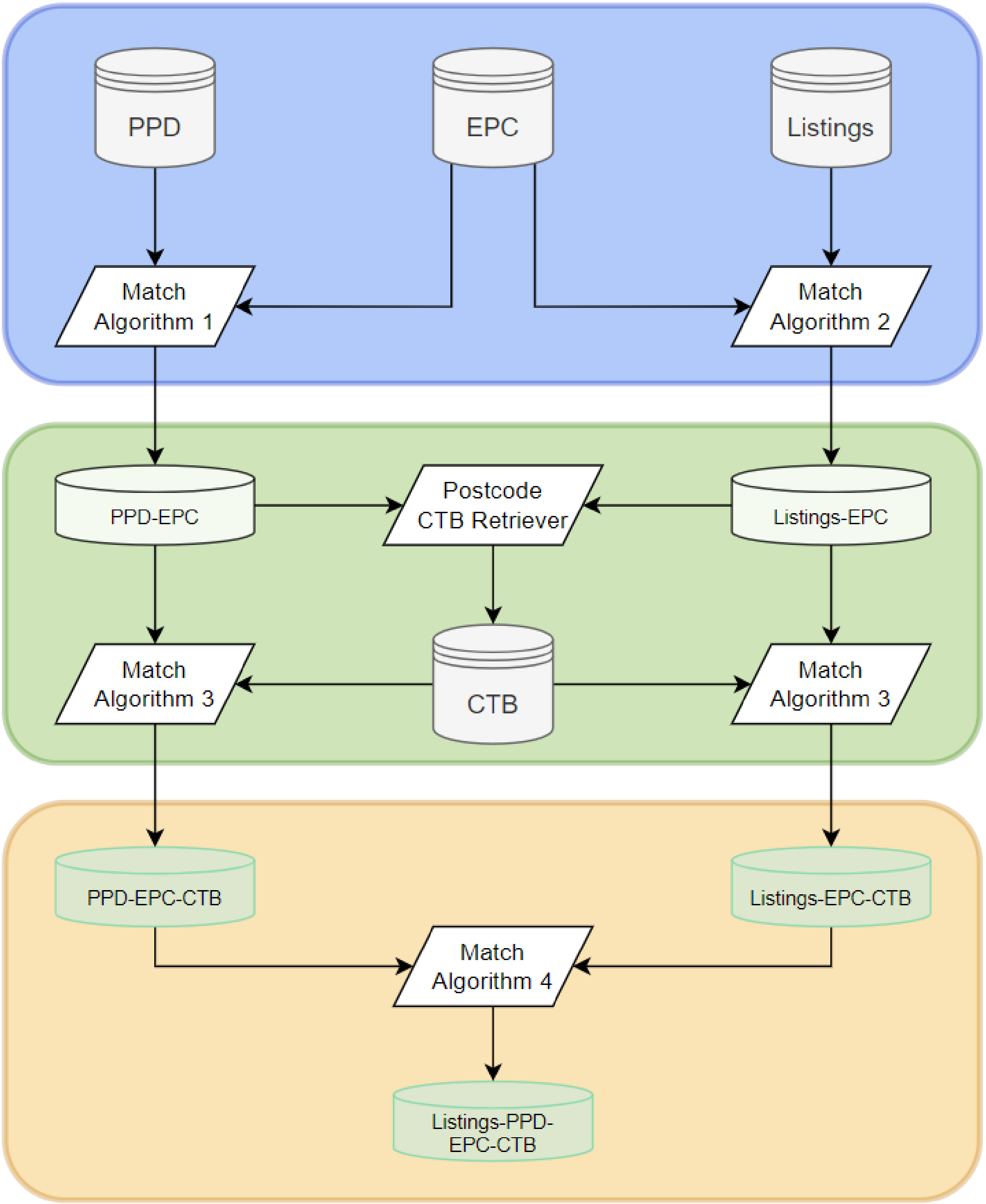


Figure 2.5: Data Integration Process

Note: This flowchart illustrates the process of data integration. In the initial stage, the EPC data is merged with the PPD and Listings datasets separately. In the subsequent stage, the CTB is collected and integrated into the two matched results, resulting in the creation of two new datasets - PPD-EPC-CTB and Listings-EPC-CTB. In the final stage, the two resulting datasets are merged to generate a comprehensive data set - Listings-PPD-EPC-CTB.

EPC data had the same postcode and the same address-associated number, it was almost certain that the match was accurate.

To further refine the matched results and ensure their accuracy and consistency, additional information from both datasets was utilised. For example, the matched records should have the same property type, and the energy efficiency assessment inspection date in the EPC data should be a date prior to the transfer date in the PPD due to legal requirements.

Finally, the PPD-EPC matched results were obtained by including only the best matches in the EPC data for each transaction in the PPD. The best matches were those with perfectly matched address strings or identical address-associated numbers. If one transaction in the PPD had multiple linked records with a perfect match in addresses in the EPC data, only the record with the inspection date closest to and prior to the transfer date was retained in the results.

From January 1, 2018 to March 31, 2021, there were 2,500,895 residential properties transferred with a full market value that were registered with the Land Registry in England. The matching algorithm identified matched records for 2,389,692 out of these transactions from EPC data, which represents a success rate of 95.6%. This demonstrates the effectiveness of the algorithm in accurately linking records from different sources.

Listings Data - Energy Performance Certificates Data

Similarly, I develop another algorithm (Match Algorithm 2 in Figure 2.5) to link the records from the listings data with the records in the EPC data. The algorithm first utilised the UPRN as the common identifier. If the UPRN was not available or did not produce a match, the algorithm then used detailed address information and postcodes as shared identifiers, which has a similar address matching mechanism to the PPD-EPC data matching algorithm. This approach allowed for accurate and consistent matching of records between the listings data and EPC data.

Council Tax Band Data - Energy Performance Certificates Data

In the second phase of the data integration process, I develop another algorithm (Match Algorithm 3 in Figure 2.5) to link records from the CTB data with the records in the EPC data. It was accomplished using detailed address information and postcodes as shared identifiers. To optimise the data integration process, only postcodes that exist in PPD-EPC matched result or Listing-EPC matched result are used to retrieve the CTB through the CTB web-scraping program (Postcode CTB Retriever in Figure 2.5).

This algorithm has a similar address matching mechanism to the PPD-EPC data matching algorithm. In addition, another function was introduced in this algorithm to obtain the CTB for each linked result in the previous phases as accurately as possible. If a record in the EPC data did not have a perfect match result by address in the CTB data but had multiple fuzzy matched results with the same postcode and they all shared the same CTB, it was considered to be a correct CTB for the record in the EPC data. This allowed for the inclusion of as much CTB data as possible without sacrificing the accuracy of the results.

Finally, the CTB was integrated into the two big datasets created in phase one to produce the PPD-EPC-CTB and Listing-EPC-CTB datasets.

PPD-Listing-EPC-CTB Data

During the final phase of the data integration process, the PPD-EPC-CTB and Listing-EPC-CTB datasets were linked together by another algorithm (Match Algorithm 4 in Figure 2.5) that mainly uses the "LMK_KEY" variable exits in EPC as the unique identifier between the two datasets. Potential mismatches are eliminated by cross-referencing all information in the matched results in this algorithm. For instance, mismatches were discarded if the transfer date occurred before the listing date, or if there were inconsistencies in property type, built form, or number of rooms from different data sources. Another example of a potential mismatch was if the transaction price was substantially different from the final asking price. For property transactions with multiple matched listings, the earliest one was kept in the big data. In cases where properties were briefly taken off the market and then relisted again, only the earliest listing before the corresponding transaction was included in the data for the purpose of recording the true exposure time of the property on the market. The resulting big datasets can then be utilised for a wide range of purposes, including real estate market analysis and policy making.

2.3 Other Variables in Big Data Sets

This section introduces several variables of significant interest and importance, which were not covered in previous sections.

Chain-free

The listings data provides us with information that can be used to determine whether the seller is "chain-free." A chain-free seller has no other transactions dependent on the sale of the property, and this information can provide insight into the seller's financial constraints. Chain-free sellers may be less financially constrained than others, and this information can help buyers in their decision-making process.

A property chain refers to a sequence of transactions that must take place for a property to be sold. It involves multiple buyers and sellers, each dependent on the sale or purchase of another property up or down the chain. For example, a person selling a property and wanting to buy another is financially dependent on the sale of their current property to proceed with the purchase of a new one. The person they sell their property to is also financially dependent on the sale of their current property, and so on, creating a chain of buyers and sellers. The longer the chain, the more complex and time-consuming the process becomes.

In contrast, the term "chain-free" in an advertisement indicates that the seller is not dependent on the sale of their current property to proceed with their next steps, which can include buying another property or not. This can make the process of buying or selling a property faster and less complicated, as it eliminates the need to coordinate the sale of multiple properties simultaneously. Properties advertised as "chain-free" are often more appealing to buyers, as they allow for a more flexible and efficient buying process. Therefore, if a property is described as "chain-free," it can be viewed as an indicator that the seller is not financially constrained, and the buying process is likely to be more streamlined.

Price modifier

The "price modifier" is an essential variable provided by the listings data, encapsulating the strategy and motivation of sellers and agents, and is a critical control variable in all models in this study. Five common listing strategies observed in the UK market include fixed price, guide price, offers around, offers over, and price on request.

"Fixed Price" is a pricing strategy in which sellers indicate a specific and non-negotiable price at which they hope to sell their property. This approach sets an upper limit on the offers that will be received, and while the property may occasionally sell for more than the fixed price, it is less likely. The advantage of this strategy is that it can lead to a quick sale, but it may also reduce the potential for competition among buyers and decrease their willingness to make higher offers.

A "Guide Price" is an estimated range or value provided by the seller or estate agent for a property that is being offered for sale. It serves as an invitation to make an opening offer or to negotiate. The guide price is not a fixed or final price, and the seller and estate agent may not be entirely certain about the property's value. They may use guide prices as a way to gauge interest in the property and to fine-tune the price based on offers received. Alternatively, a guide price may be used when there is disagreement about the property's market value, and the seller hopes to achieve a sale that is higher than the estate agent's valuation.

"Offers Around" (or "Offers in the region of") is a pricing strategy used in property advertising in the UK that falls between the Fixed Price and Offers Over methods. This strategy is employed when the seller and agent are uncertain about the appropriate price for the property, and it allows the seller to set a closing date and consider multiple offers if there is significant interest in the property.

”Offers Over” (or ”Offers in excess of”) is frequently employed as a means of promoting a property at a price slightly below its estimated market value. The objective of this strategy is to generate interest and potentially initiate a competitive bidding process among prospective buyers. The seller establishes a minimum asking price, which they expect offers above, to give potential buyers an indication of the minimum amount that should be offered to secure the property.

”Price on request” (or ”Price on Application”) is a pricing strategy where the seller only reveals the price of a property to serious buyers who have shown interest in the property. This approach is often used for luxury properties or unique properties that are difficult to price. However, in 2022, The National Trading Standards Estate Agents and Letting Agents Team has prohibited the use of ”Price on request” in property listings, citing it as misleading and in violation of consumer protection legislation. This decision is expected to enhance transparency in the housing market by eliminating the use of ”Price on request” and ensuring clearer information on property asking prices.

2.4 Summary

Research on the UK residential housing market has been ongoing since the 1970s, with aggregate datasets such as building society mortgage data, and local estate agent survey data heavily utilised. However, these datasets have limitations, including a lack of detailed information and potential biases from their limited sample size. Although the availability of the Land Registry PPD as open data since 2013 has been transformative, its significant shortcomings include the absence of physical property characteristics such as floor size. The current research on the residential market in the UK is hindered by the lack of a comprehensive property data set containing transaction, listings information and property attributes, which underscores the need for flexible and comprehensive data collection methods and frameworks.

This chapter aims to bridge the gap in constructing big datasets on housing by leveraging several openly accessible data sources. To achieve this, a series of programs and algorithms have been developed to construct several property-level big datasets. The chapter starts by offering a comprehensive overview of the data sources used to construct large datasets in residential housing. Next, it outlines the data linking process and provides a detailed description of the algorithms used in constructing these datasets. Lastly, several variables that are of significant interest and importance in this study are introduced. These datasets are among the most comprehensive used in similar studies and offer unique insights into the UK residential market.

In the context of this thesis, the proposed data integration framework effectively addresses the standard Big Data dimensions of volume, velocity, and variety.

Notably, the dataset encompasses transactions across a wide spectrum, ranging from the smallest sample, representing one-third of the population transactions, to the largest sample, accounting for a substantial 92%. This diversity in volume underscores the framework's scalability and adaptability to handle datasets of varying sizes.

The framework's monthly data scraping process plays a pivotal role in tackling velocity concerns. By integrating a dynamic and timely data acquisition strategy, the framework ensures that the latest property listings' information is promptly incorporated, which supports the development of insights into the current market.

In terms of variety, the framework employs text analysis techniques to extract valuable information from property descriptions within advertisements. This includes the creation of new variables for modelling, such as indicator variables like "chain-free," "garden," and "garage." This not only showcases the framework's adaptability to diverse data sources but also emphasises its ability to enrich datasets with nuanced variables derived from unstructured text.

Moreover, as a forward-looking initiative, the framework lays the groundwork for future investigations by indicating the potential integration of image analysis. This would involve extracting insights from images within advertisements, collected through the existing data scraping program. Incorporating image analysis represents an avenue for expanding the variety dimension, offering new possibilities for understanding and modelling real estate data.

Descriptions for Variables Used in This Research

Table 2.2: Variable Description

Variable	Description
Transaction price	Sale price stated on the transfer deed.
Initial listing price	The price on a property listing appeared on the Zoopla at the first time.
Price spreads	Final listing price - Transaction price.
TOM	Time on the market (in days).
RUC	Rural-Urban Classification: Rural, Urban City&Town, Urban Conurbation.
Property type	House vs. Flat.
Duration	An indicator of duration of ownership that is either Freehold or Leasehold.
New built	Yes or No (Y or N).
Price modifier	Fixed price, Guide price, Offers around, Offers over or Price on request.
Chainfree	Indicator for whether the property listed is chain free, which means the property you want to buy is not reliant on the successful purchase or sale of other properties. As a chain-free buyer, your purchase is not dependent on the sale of a property you currently own. For example, all first-time buyers are chain-free buyers.
Garage	Indicator for presence of a garage.
Driveway	Indicator for presence of a driveway.
Garden	Indicator for presence of a garden.
Total floor area	Total floor area in Square Meter.
Final listing price	The final revised listing price.
Num habitable rooms	Number of any living, dining room, bedroom, study and similar. Excluded from the room count are any room used solely as a kitchen, utility room, bathroom, cloakroom, en-suite accommodation and similar and any hallway, stairs or landing; and also any room not having a window.
Num Extension	The number of extensions added to the property.
Num open fireplaces	Number of open fireplaces in the property.
Current energy efficiency	Displayed on EPC. Based on cost of energy multiplied by fuel costs. (£/m2/year where cost is derived from kWh)
Potential energy efficiency	Displayed on EPC. The potential energy efficiency rating of the property.
Environment impact current	Displayed on EPC. The Environmental Impact Rating. A measure of the property's current impact on the environment in terms of CO2 emissions. The higher the rating the lower the CO2 emissions (in tonnes/year).
Environment impact potential	Displayed on EPC. The potential Environmental Impact Rating. A measure of the property's potential impact on the environment in terms of CO2 emissions after improvements have been carried out. The higher the rating the lower the CO2 emissions (in tonnes/year).
Energy consumption current	Displayed on EPC. Current estimated total energy consumption for the property in a 12 month period (kWh/m2). Displayed on EPC as the current primary energy use per square meter of floor area.
Energy consumption potential	Displayed on EPC. Estimated potential total energy consumption for the Property in a 12 month period. Value is Kilowatt Hours per Square Meter (kWh/m^2).
CTB	Council Tax Band, from A to H, depending on the price they would have sold for in April 1991, assessed by Valuation Office Agency (VOA).

List of Main Algorithms/Program Developed for this Research

Table 2.3: List of Main Programs/Algorithms

Program/Algorithm	Description
Monthly listings collector	Collects information on property listings advertised on the market through Zoopla's API in England. This program runs on a monthly basis to ensure that the collected data is up-to-date.
Listings address retriever	A web-scraping program that collects detailed addresses for the listings collected from Zoopla's API.
Listings UPRN retriever	A web-scraping program that collects UPRNs for the listings collected from Zoopla's API.
CTB data retriever	A web-scraping program that collects council tax band information for each postcode area from the government's lookup portal.
EPC data preprocessor	Extracts and combines EPC records in England. Preprocesses address information for each record in EPC data for matching purposes. Fills missing data based on corresponding text description with text mining method.
Listings data preprocessor	Preprocesses address information for each record in listings data for matching purposes. Creates new variables based on the listing's description (e.g. chain-free, garden, driveway, garage).
PPD data preprocessor	Preprocesses address information for each record in PPD data for matching purposes. Retains records of sales for full market value in England.
Match Algorithm 1	An interactive matching program for the integration of PPD and EPC data.
Match Algorithm 2	An interactive matching program for the integration of Listings and EPC data.
Match Algorithm 3	An interactive matching program for the integration of CTB into other datasets.
Match Algorithm 4	A matching program for the merging of PPD-EPC-CTB and Listings-EPC-CTB datasets.

Chapter 3

The Price–Time-on-Market Puzzle Revisited

3.1 Introduction

The residential real estate market has been widely studied with regards to transaction processes, but the microstructure of the market remains largely unknown. Housing plays a significant role in household portfolios, but its heterogeneity and illiquidity pose challenges for market participants in determining the true value of residential transactions. The standard hedonic model proposed by Rosen (1974) suggests that properties are valued for their utility-bearing attributes, including physical characteristics and location-related amenities and services. However, even when controlling for these attributes, prices remain dispersed in the local market (He et al. 2017).

Sellers may set prices significantly different from the market value of their properties, either by overpricing or underpricing. In turn, buyers may face the classic lemon problem, which is a situation where buyers have difficulty assessing the quality of a product or service due to the asymmetry of information between the buyer and the seller. Both parties may base their valuations on property characteristics and past transaction prices in the neighbourhood. This can offer insights into the market value of similar properties and contribute to the valuation of the property in question. However, it is important to note that this is not a foolproof method. Properties are heterogeneous and have unique characteristics, making it difficult to compare them and accurately determine their true value. Additionally, past transaction prices in the neighbourhood may not necessarily reflect the current market conditions, which can lead to inaccurate valuations. The greater the discrepancy between the buyer's and seller's valuations, the longer the duration for completing the transaction, known as the time on the market (TOM). Once the buyer's and seller's valuations align, the value and TOM are established, and a transaction takes place. The duration of this process can extend for several months and is influenced by factors such as the initial asking price, the level of bargaining, and institutional parameters such as legal requirements for title and deed.

The search for the true value of a property is formalised in the search and matching model of housing markets. This process ultimately leads to the attainment of equilibrium market value. The model acknowledges that buyers and sellers do not possess perfect information about the market and each other, which leads to a search process. Additionally, it recognises that the value of a property is not fixed, but rather, it is determined through the negotiation process between buyers and sellers. This provides a fundamental explanation for the violation of the law of one price in the housing market. Furthermore, the model demonstrates that the search duration, or TOM, is a critical determinant of a property's transaction price (TP). The negotiation process between buyers and sellers implies that both variables are jointly determined. Based on search theory (Anglin et al. 2003; Krainer and LeRoy 2002; Wheaton 1990), the TP and TOM are dependent on the probability of sale. Therefore, a positive relationship between TP and TOM is expected (Hayunga and Pace 2019).

Despite theoretical studies suggesting a positive relationship between price and TOM, the dependence is less clear in empirical studies. The effect of TOM on the TP and the effect of the TP on TOM have been the subject of extensive research, with several studies providing a comprehensive survey of the literature on the price-TOM relationship. For example, Johnson et al. (2007) examined the probability of sale, and Benefield et al. (2014) summarised simultaneous modelling techniques. However, the results of these studies are inconsistent. Among 429 estimations, price and TOM were positively significantly correlated in 111 instances, negatively significantly correlated in 176 cases, and had an insignificant correlation in 142 instances (Benefield et al. 2014). The discrepancy in results may be due to differences in data and sample periods, but as He et al. (2017) argues, the empirical evidence is so divided that it defies any reasonable explanation.

One of the main challenges in explaining the relationship between (TP) and TOM is the endogeneity between the two variables, as well as the fact that they are jointly determined. To accurately identify this relationship, a transaction-level database was constructed by merging various sets of data. As mentioned in Chapter 2, big data techniques were employed to collect information from property listings between January 2018 and March 2021 in England, and this data was merged with data from the PPD, EPC data, and CTB data, creating a rich micro dataset. The final dataset covers approximately one-third of residential properties sold for full market value that were lodged with the HM Land Registry in England during the specified time frame. To the best of my knowledge, this is one of the largest datasets used in similar studies, providing a robust and comprehensive examination of the relationship between TP and TOM.

The dataset includes unique variables that have not been previously used, aiding in identifying the relationship between transaction price (TP) and time on the market (TOM). Additionally, many other property attributes have been measured by a third party under the same standard across the nation (from the Energy Performance Certificate (EPC) data), providing high-quality control variables. Along with the initial listing date and price, various iterations of the listing price over time are also observed. This enables the observation of the duration of price revisions, allowing an assessment of how sellers were adjusting their valuation of the property. The data also contains information about the council tax band (CTB) in which the property falls, which is an important property value indicator.

In addition, textual analysis was employed to extract a variable called 'chain-free'[1]

[1] In the context of property listings, the term "chain-free" refers to a property that is available

from the description in the property's advertisement. It indicates whether the seller would buy a new home conditioned on the sale of the current place. This is often the case when the purchase of a new home for the seller is associated with using the proceeds from the sale to pay for the down payment of the new home. This variable can be seen as a measure of whether the seller is financially constrained.

This chapter makes several contributions through the use of the big dataset. Firstly, a novel method for measuring overpricing is proposed, following the methodology of Knight (2002) and Anglin et al. (2003). Overpricing is measured as an initial markup larger than five percent, which allows for negotiation frictions that are not typically considered overpricing in practice. The markup is the difference between the initial listing price and the TP in percentages. Omitting this variable can lead to biased results on the relationship between TP and TOM, as demonstrated in the findings of this chapter. This supports the theoretical finding in Taylor (1999) that overpricing can potentially lead to a negative price-TOM relationship.

Secondly, given the simultaneous nature of the relationship between price and TOM, ordinary least squares (OLS) estimations may lead to biased results. However, according to Benefield et al. (2014), 40 out of 68 empirical papers have used OLS to model this relationship. To address this endogenous relationship, a simultaneous equations model (SEM) is used. Two novel instrumental variables (IVs) are employed in a two-stage least squares (2SLS) regression framework, which have not been used previously due to the limitation of datasets.

As mentioned in Chapter 2, the CTB reflects a property's valuation as of 1 April 1991, which is assessed by the Valuation Office Agency based on the same criteria across the nation. Local authorities also decide the council tax paid for each band; the higher the band, the more expensive the council tax. Since all properties are categorised into bands based on the same criteria, the CTB is exogenous to all par-

for purchase without the need for the buyer to sell an existing property first. This means that the seller is not part of a chain of buyers and sellers, and the transaction can proceed without any delays caused by the need for other properties in the chain to be sold first. This can make the process of buying or selling a property faster and less complicated as it eliminates the need to coordinate the sale of multiple properties at the same time. A property that is described as "chain-free" is often more appealing to buyers as it allows for a more flexible and efficient buying process.

ticipants - buyers, sellers, and agents - involved in the transaction process. Therefore, the CTB is a high-quality IV for the TP. In addition, the duration of price revisions is utilised as an instrumental variable for TOM. This approach is based on the following logic[2]: a seller sets an initial asking price but may revise it while waiting for potential buyers. The revision duration is the period from the date of the initial listing to the date of the final listing price revision. Essentially, the final listing price directly affects the TP and the remaining market duration. However, the revision duration only affects the TOM and not the price, as long as the quality of the property is controlled for, which can effectively eliminate the potential for a stigma effect of a lemon property.

The results of this chapter present evidence that the price-TOM relationship is positive and simultaneously determined, which is in line with the search theory. It is suggested that previous studies that have found the opposite sign may be due to not accounting for sellers' overpricing and a lack of suitable IVs due to data limitations.

Moreover, it is indicated that chain-free sellers who are not constrained by the need to sell their current property before buying a new one tend to set lower initial asking prices and accept lower transaction prices, all else equal. This finding may be attributed to the presence of agency costs.

The structure of this chapter is as follows: In the next section, a review of the most relevant studies that analyse the relationship between price and TOM is presented. The following section provides an overview of the data used in this chapter, including descriptive statistics. The fourth and fifth sections present the methodology employed and the results of the empirical analysis, respectively. Finally, a conclusion is provided.

[2] More details are discussed in Section 3.4.2

3.2 Literature Review

3.2.1 Research Before 2014

Prior to 2014, the relationship between TOM and price in the housing market appeared in a substantial amount of research. For various research purposes, the price-TOM relationship (which includes TOM as an independent variable in property price estimation or TP as an independent variable in TOM estimation) has been examined, often as a by-product of investigating the main research question. A notable consistency from these studies is the inconsistent nature of the relationship between price and marketing time. Johnson et al. (2007) and Benefield et al. (2014) comprehensively presented the inconsistent empirical estimations of the price-TOM relationship in studies before 2014. Therefore, the following discussion mainly focuses on the most recent studies.

Studies demonstrating an inverse association between prices and TOM frequently attribute it to overpricing or stigma effects associated with structural defects in the property. Taylor (1999) posits that overpricing could lead to a vicious cycle of rejected offers and reductions in the asking price, resulting in the property being sold at a discounted price or withdrawn from the market altogether. An overpriced property may require one or more price reductions to induce a transaction, which increases TOM and potentially leads to a negative price-TOM relation. While this relationship can be tested empirically, data limitations make it challenging to measure the level of overpricing accurately. Knight (2002) presents empirical evidence that initial mispricing costs sellers more time and money, and houses with substantial listing price changes have longer TOM and ultimately sell at lower prices. Anglin et al. (2003) introduces a variable called the degree of overpricing (DOP), which measures the difference between the actual listing price and the expected listing price, and shows that increases in DOP increase TOM.

The stigma effect hypothesis suggests that TOM serves as an indicator of a property's quality, similar to the asking price (Taylor 1999). The idea is that if a prospective buyer has access to the initial listing date, a property without any hidden defects would likely have been sold before the buyer had the opportunity to view it. Therefore, a longer TOM may indicate the presence of a defect that is not immediately apparent to the new buyer but was discovered by previous viewers. Additionally, even if there is no hidden defect, TOM may be extended due to prospective buyers being suspicious of the property, resulting in an opposite of a herding effect.

3.2.2 Recent Studies

Since 2016, several studies have been conducted to better understand the relationship between price and TOM. Since these two variables are jointly determined, the error terms associated with the TOM and TP models are correlated, which can lead to biased and inconsistent results when using OLS estimation. Despite this, a significant number of studies, approximately 40 out of the 68 investigated by Benefield et al. (2014), still use OLS models.

Dubé and Legros (2016) propose a 2SLS approach with instrumental variables constructed from information from previous neighbourhood transactions, which are argued to be exogenous to price and TOM. This approach is based on the idea that past transactions in the vicinity of a property are exogenous to its current sale, and as such, the distance-weighted average neighbourhood TOM and TP are proper instrumental variables of the current TOM and TP. Because of the unidirectional temporal property, the past can influence the future, but the inverse is not true. Utilising a dataset containing 29,471 transactions in the suburban neighbourhood of Montreal from 1992 to 2000, they find that, everything else being equal, TOM is negatively related to the final sale price.

However, McGreal et al. (2016) directly investigated the spatial dependence in TOM for residential properties. They found that neighbourhood TOMs are randomly distributed with no significant correlation, and this finding is consistent over time. Besides, their models do not control for overpricing, which could lead to biased estimations as previously mentioned.

He et al. (2017) recently adopted the concept of the stigma effect as one of the two forces that causes an inverted U-shaped relationship between price and TOM. It arises from two opposing effects of TOM on the TP. On the one hand, there is an exposure effect in which a longer TOM increases the seller's probability of encountering a higher offer from buyers. On the other hand, there is a stigma effect in which a longer TOM might signal possible hidden defects of the property. Using a sample of 158,288 single-family home sales from Virginia, their model added a square-TOM term in a 2SLS estimation and found a positive coefficient (around 0.002) for the TOM term and a negative coefficient (around -0.000005) for the square of TOM, which implies an inverted U-shaped relationship between the TP and TOM.

However, Hayunga and Pace (2019) includes a measure of structural quality from Genesove and Mayer (2001) and finds no empirical evidence to support the stigmatisation hypothesis. They argue that the stigma effect may be possible at an individual property level, it should be idiosyncratic for a few lemon[3] properties and not systematic across the market. Similarly, their models do not control for overpricing, which could lead to biased estimations as previously mentioned.

Hayunga and Pace (2019) conducted a study based on survey data, comprising over 3,100 observations, to investigate the discrepancy between theoretical and empirical results regarding the relationship between price and TOM. Their study primarily focuses on the quality of instrumental variables. They created a statistically strong instrumented TOM by incorporating variables[4] from survey data in the first stage of the 2SLS estimation process and demonstrated a positive correlation between price and TOM. They compared this result to another model with weak instrumented TOM and suggested that the weak instrumental variables are responsible for inconsistent empirical relationships when modelling the price-TOM relationship.

Following similar U-shaped specifications as presented in He et al. (2017) in the price model, with a square-TOM term, they found breakpoints at 52 weeks for the non-transformed TOM model and 65 weeks for the log-transformed TOM model;

[3] In American slang, a 'lemon' refers to a car that is discovered to be defective after its purchase. Akerlof (1970) conducted a widely cited seminal study that delves into the concept of 'lemons' within the context of asymmetric information in markets.

[4] These variables, coded as binary indicators, measure search costs for selling to friends or acquaintances and additional marketing methods, including open houses and various advertising channels such as magazine, flyer, print, and television advertising.

however, these were not statistically significant.

3.2.3 Research related to TOM or Price

Several studies in the literature are closely related to the Price-TOM puzzle. These studies provide valuable guidance in selecting appropriate control variables and constructing instruments for TOM.

The initial listing price plays a crucial role in the negotiation process from the outset. Homeowners who anticipate incurring a loss tend to set a higher listing price initially but eventually attain higher selling prices with a much lower probability of sale failure than other sellers (Genesove and Mayer 2001). Moreover, Levitt and Syverson (2008) demonstrate a principal-agent problem, whereby realtors have an incentive to induce clients to sell cheaply and quickly, which, in turn, shortens the TOM.

Changes in the listing price can also significantly impact the transaction process. Lazear (1986) presents a theory of pricing behaviour, where price revision is allowed over time, and product demand is uncertain. According to this theory, the seller can learn about the buyer's valuation through a function of time that describes the initial listing price and its changes. Building on this theory, Knight (2002) argues that sellers should set a lower initial listing price and not change it when the number of customers is small. This is because the seller can learn little about the buyer's valuation during the first period. However, in a more actively traded market, a relatively higher initial listing price with subsequent changes can provide more knowledge of the distribution of buyers' valuations. Loss-averse sellers are more likely to revise their listing prices downward more aggressively than others (Liu and Vlist 2019). Similarly, Wit and Klaauw (2013) estimated the causal effect of lowering the listing price on TOM and found that listing price reductions significantly increase both the house selling rate and its withdrawal rate.

Pricing strategy is another factor that affects the transaction process. Cardella and Seiler (2016) investigated the effect of rounded, just below, and precise listing prices on agents' behaviour and found that a highly precise price led to the highest transaction price with the smallest discount on the listing price. In contrast, just below pricing generated the lowest final price with the largest discount. Similarly, Allen et al. (2005) found that the use of a range pricing strategy led to longer TOM for properties but did not significantly impact the TP.

3.3 Data and Variables

As highlighted in the literature, this study incorporates a variable known as the "price modifier," which captures the strategies and motivations of sellers and agents and serves as a crucial control in modelling transaction behaviour. In the UK housing market, five commonly used listing strategies are fixed price, guide price, offers around, offers over, and price on request. The utilisation of these pricing strategies can vary depending on the unique factors involved in the selling process. A "fixed price" strategy establishes an upper limit on the offers received, resulting in a quicker sale but with a lower potential sale price. On the other hand, a "guide price" is an estimated value that serves as an invitation to make an opening offer or negotiate, and can be used to gauge interest in the property or when there is a disagreement about the market value. When the seller and agent are uncertain of the appropriate price, the use of "offers around" is employed, allowing for negotiation and consideration of multiple offers. Conversely, "offers over" is used to generate interest and potentially initiate a bidding war, with the seller setting a minimum asking price. Additionally, "price on request" is a pricing strategy where the seller only reveals the price to serious buyers who have shown interest in the property.

In this research, a new variable called "Price Revise Times" was created to provide information about the seller's characteristics. If the seller is motivated or if there is an overpricing, the price tends to be revised more frequently. Additionally, a variable called "Price Revise Duration" was created to measure the time a seller took to revise the asking price. It is calculated as the difference between the initial published date and the date when the final revised price was updated on the advertisement.

In addition to the variables discussed in Chapter 2, variables related to regional and local authorities were also created for each property based on its postcode and address. These variables were utilised for robust tests and as location-related fixed effect controls. During the matching process outlined in Chapter 2, properties that were listed under multiple listing IDs in Zoopla's listing data were identified. In instances where properties were briefly taken off the market and then relisted, only the earliest listing prior to the corresponding transaction was included in the data in order to accurately record the property's true exposure time on the market. As a result, the TOM in our sample may be slightly larger than the numbers commonly reported by agents' websites.

3.3.1 Data Summary

Table 3.1: Summary of Data Size by Year

	2018	2019	2020	2021	Total
Population	800, 258	768, 855	673, 257	258, 525	2, 500, 895
Sample	234, 617	266, 459	251, 165	96, 052	848, 293
Sample/Population	0.293	0.347	0.373	0.372	0.339

According to the PPD published in December 2021, a total of 2,500,895 residential properties were sold at full market value in England from January 2018 to March 2021. Through the process of cleaning our sample data by removing observations with missing values and trimming dependent variables by 0.1% to exclude outliers, the final sample contains 848,293 observations. This accounts for approximately 34% of the transaction data population. It is noteworthy that the observations in the sample from 2018 constitute a slightly smaller proportion of the population in comparison to the sample size in other years, as the listing data were collected from January 2018, and some of the properties that were transacted in 2018 were listed on the market prior to 2018. A robustness test is presented at the end of the results section to demonstrate that our models can generate consistent estimates even in the presence of potential sampling bias in the data.

Table 3.2 and 3.3 report the summary statistics of numerical and categorical variables in the final cleaned data. The numerical variables are described by their minimum, 25th percentile, mean, median, 75th percentile and maximum. The categorical variables are described by their count and fraction. The descriptions of all variables are listed in Table 2.2 in Chapter 2.

Table 3.2: Summary Statistics of Numerical Variables

Statistic	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
TOM	1	120	220.70	171	262	1,164
TP	15,000	172,000	310,574.00	257,500	382,500	2,050,000
Initial Listing Price	5,000	179,950	326,217.00	270,000	400,000	3,750,000
Price Revise Times	0	0	0.52	0	1	6
Price Revise Duration	0	0	31.85	0	35	1,684
Total Floor Area	26	73	98.64	89	114	446
Num Habitable Rooms	1	4	4.87	5	6	11
Num Open Fireplaces	0	0	0.15	0	0	40
Num Extension	0	0	0.61	0	1	4
Current Energy Efficiency	1	57	62.87	64	70	142
Potential Energy Efficiency	1	77	80.68	82	85	142
Environment Impact Current	1	50	59.03	60	68	136
Environment Impact Potential	1	73	77.91	80	84	139
Energy Consumption Current	−257	193	256.30	242	303	1,831
Energy Consumption Potential	−338	88	127.90	114	152	1,417

Observations: 848,293

The average TP for properties in the dataset is £310,574, while the median is £257,500. The positive skewness of this variable suggests that the majority of observations have a relatively low sales price, with a few observations having a significantly higher TP. The minimum and maximum TP in the dataset are £15,000 and £2,050,000, respectively. The average and median initial listing prices for the properties are £326,217 and £270,000, respectively. These values are slightly higher than the corresponding statistics for the TP. This could be attributed to the fact that sellers often set their initial listing price higher than what they expect to receive, as it allows for more flexibility in the negotiation process, which is a common practice in the housing market.

On average, it takes 220.7 days to sell a property in the sample period. The median TOM is 171 days, which is approximately 2 months shorter than the mean TOM. The 75th percentile of TOM is 262 days, indicating that most properties are sold within 9 months on the market. The maximum TOM is 1,164 days, which is approximately 3.2 years.

The "Price Revise Times" variable represents the number of times the price was revised before the property was sold, with a mean value of 0.52 and a median value of 0. This variable has a positive skewness, suggesting that the majority of properties were sold without any revisions to the price, while a small proportion of properties had multiple revisions before they were sold. The "Price Revise Duration" variable represents the amount of time, in days, that the seller spent revising the asking price on their advertisement. On average, sellers spent 31.85 days revising the asking price of their properties. The median value of this variable is 0, indicating that the majority of properties were sold without any revisions to the asking price. The minimum and maximum values of the variable are 0 and 1,684 days, respectively.

The "Total Floor Area" variable represents the total area of the property measured in square meters. The mean value of this variable is 98.64 square meters, while the median value is 89 square meters. The variable has a relatively normal distribution, with a minimum total floor area of 26 square meters and a maximum total floor area of 446 square meters. The "Num Habitable Rooms" variable represents the number of rooms in the property that can be used as living spaces. On average, properties in the dataset have 4.87 habitable rooms, with a median value of 5. The majority of properties in the dataset have at least 4 habitable rooms. The "Num Open Fireplaces" variable represents the number of open fireplaces in the property. On average, properties in the dataset have 0.15 open fireplaces, with a median value of 0. The majority of properties do not have any open fireplaces. The "Num Extension" variable represents the number of extensions in the property. On average, properties in the dataset have 0.61 extensions, with a median value of 0. The majority of properties in the dataset do not have any extensions.

The average "Current Energy Efficiency" score is 62.87, while the "Potential Energy Efficiency" score is 80.68. This indicates that most properties have the potential to significantly improve their energy efficiency. Similarly, the average current environment impact score is 59.03, while the mean potential environment impact is 77.91.

This means that most properties have the potential to reduce their impact on the environment. Likewise, the majority of properties can significantly reduce their energy consumption. The average current energy consumption is 256.30, while the potential mean is 127.90.

Table 3.3 presents an overview of the property transactions in the sample. The majority of the transactions (79.5%) are for houses, while only 7.5% are for apartments. Additionally, 88.2% of the properties in the sample have a garden. 47.5% of properties have a garage, while 41.8% have a driveway. Most transactions in the sample fall under council tax band B, C, or D, which accounts for over 64% of the sample. Approximately 32% of the transactions are listed as "chain-free." Only 3.8% of the properties are listed as "Fixed Price." The majority of the properties (73.6%) are listed as "Guide Price", while 18.5% are listed as "Offers Over." In terms of ownership, 86.3% of the properties are freehold, while 13.7% are leasehold. It is worth noting that newly built properties only account for 0.15% of the transactions in our sample. This can be attributed to several factors, including a limited supply of new properties on the market, a lack of advertising for some new properties on Zoopla, and a failure to collect detailed addresses for matching for those new properties that are advertised on Zoopla.

3.4 Methodology

In this study, I adopt the standard simultaneous equations model (SEM) as the fundamental framework for the jointly determined price and TOM trade-off. The fundamental SEM follows Dubé and Legros (2016) and is given as:

$$log(TOM) = \beta_1 log(TP) + \boldsymbol{\beta_2} \boldsymbol{X} + \boldsymbol{\xi}, \tag{3.1}$$

$$log(TP) = \alpha_1 log(TOM) + \boldsymbol{\alpha_2} \boldsymbol{X} + \boldsymbol{\varepsilon}. \tag{3.2}$$

Both equations abide by the hedonic demand theory, which posits that properties are valued for their utility-bearing attributes. The matrix $\boldsymbol{X}$ represents a set of control

Table 3.3: Summary Statistics of Categorical Variable

Variable	Value	Count	Fraction
Chain-free	Yes	270, 877	0.3193
	No	577, 416	0.6806
Garden	Yes	748, 505	0.8824
	No	99, 788	0.1176
Garage	Yes	402, 494	0.4745
	No	445, 799	0.5255
Driveway	Yes	354, 283	0.4176
	No	494, 010	0.5824
Council Tax Band	A	109, 870	0.1295
	B	163, 754	0.1930
	C	209, 712	0.2472
	D	169, 722	0.2001
	E	110, 555	0.1303
	F	54, 406	0.0641
	G	28, 842	0.0340
	H	1, 432	0.0017
Price Modifier	Fixed Price	3, 263	0.0038
	Guide Price	624, 419	0.7361
	Offers Around	63, 030	0.0743
	Offers Over	156, 646	0.1847
	Price on Request	935	0.0011
Duration	Freehold	731, 964	0.8629
	Leasehold	116, 329	0.1371
Old New	Old	847, 058	0.9985
	New	1, 236	0.0015
Built Form	Detached	261, 988	0.3088
	End-Terrace	98, 000	0.1155
	Mid-Terrace	184, 193	0.2171
	Semi-Detached	304, 112	0.3585
Property Type	House	674, 762	0.7954
	Bungalow	101, 841	0.1200
	Flat	63, 767	0.0752
	Maisonette	7, 920	0.0093
	Park Home	3	0.0000

Observations: 848,293

variables containing property characteristics, location (Lower Tier Local Authority, LTLA), and time (both year and month) dummies to avoid any confusion between aggregate effects and estimations of the price-TOM relationship. The coefficients of interest are $\boldsymbol{\beta_1}$ and $\boldsymbol{\alpha_1}$, which indicate the effect of price on TOM and the effect of TOM on price, respectively. The vectors of coefficients of $\boldsymbol{X}$ in each equation are represented by $\boldsymbol{\beta_2}$ and $\boldsymbol{\alpha_2}$. The error terms are represented by $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$.

3.4.1 Measuring Overpricing

Overpricing in property listings can have a significant impact on both TP and TOM. It is a confounding variable when estimating the effect of TOM on price in Equation 3.2 and vice versa in Equation 3.1. The omission of an overpricing control could lead to biased estimations of both α_1 and β_1.

The negative impact of overpricing on the sale process of a property has been well-documented (Taylor 1999). However, the definition of overpricing, which is dependent on the TP, is often unclear in the real estate market where the law of one price[5] does not hold. Indirect and insubstantial evidence, such as a lack of viewings or offers for a period of time on the market, a faster sale of a neighbouring property, or a significant difference in listing price compared to a property in the vicinity, may be used to infer relevant information. As a result, measuring overpricing can prove to be a difficult task.

There have been few attempts to measure overpricing in existing research. One approach suggested by Anglin et al. (2003) is a ”degree of overpricing” measure, which is the percentage difference between the actual listing price and the expected listing price. The expected listing price is modelled using property characteristics and market variables using an OLS estimator, which is essentially a predicted selling price with a standard hedonic model. Similarly, Knight (2002) defines a listing price

[5] The law of one price, a concept rooted in economic theory, posits that identical goods or assets should sell for the same price when factors like transportation costs are not considered. The intuition behind the law of one price is based on the assumption that differences between prices are eliminated by market participants taking advantage of arbitrage opportunities. In the context of housing markets, imperfect information and negotiation processes between buyers and sellers can lead to deviations from this theoretical concept.

markup as a measure of the seller's motivation signal, which captures mispricing as follows:

$$InitialMarkup = \frac{InitialListingPrice}{TransactionPrice} - 1$$

My dataset contains both the initial listing price and the TP, which allows for a similar proxy of overpricing. The actual overpricing value would be somewhere between the initial listing price and the TP. The difference between the listing price and the TP captures not only overpricing but also other price premiums associated with negotiation frictions. To quantify overpricing, I use the initial markup following the above study, which is calculated as the difference between the initial listing price and the TP divided by the TP. To account for negotiation frictions, which are not typically considered mispricing, only markups greater than 5% are considered as measured overpricing.

$$OverPricingProxy(OPP) = \begin{cases} InitialMarkup, & InitialMarkup > 0.05 \\ 0, & InitialMarkup \leq 0.05 \end{cases}$$

If the initial markup is negative, the property is considered underpriced; if it is positive, the property is considered overpriced. The effects of underpricing and overpricing are asymmetrical and distinct. Properties experiencing underpricing tend to spend less time on the market and sell at a lower price, resulting in a positive correlation between TOM and price. However, in the case of overpricing, endogeneity concerns arise in a different way, leading to a negative relationship between price and TOM. Therefore, incorporating an overpricing control in the model specification is essential for correctly identifying the relationship between price and TOM.

3.4.2 Instrumental Variables for SEM

The joint determination of TP and TOM implies that the error terms in the TOM model are correlated with those in the TP model, and vice versa. Consequently, the use of an OLS estimator in both models can yield biased results. Despite this issue, 40 out of 68 studies investigated by Benefield et al. (2014) still employed OLS models.

To address this problem, 2SLS is a standard estimation method for controlling simultaneity. According to the rank condition for identification of structural equations, at least one instrumental variable is required for each endogenous variable, namely, the price and TOM, respectively. Notably, the IV for TOM in equation (3.1) does not enter equation (3.2), and the instrument for TP in equation (3.2) does not form part of equation (3.1).

Moreover, the effectiveness of 2SLS in controlling simultaneity is contingent on the quality of the instrumental variables. When the instruments are weakly correlated with the endogenous regressor, 2SLS is known to be biased towards the OLS estimator (Bun and Windmeijer 2011; Young 2022). Studies have shown that weak-instrument models tend to exhibit negative or insignificant slope coefficients on TOM (Hayunga and Pace 2019). Conventionally, the weak-instrument test is based on the F-statistics of the first-stage auxiliary regression, yet this test is largely uninformative of both size and bias (Young 2022). Therefore, in addition to the weak-instrument F-test, I investigate the theoretical relevance between the instrumental variables and the endogenous explanatory variables, which is often omitted in previous research due to data limitations.

In this research, I utilise CTB as an instrumental variable for the TP in the TOM equation. The CTB reflects a property's valuation as of April 1, 1991, which is assessed by the Valuation Office Agency based on the same criteria nationwide. Each year, local councils set a council tax rate for each valuation band based on the local services they provide. Notably, the council tax band, not the value, is universally assessed by the same government agency and to the same standard, rendering it exogenous to all property transaction participants, including buyers, sellers, and agents. Therefore, the CTB constitutes a high-quality instrumental variable for the TP.

To identify a suitable instrumental variable for TOM, I utilise the information contained in the seller's revision process of the listing price. Specifically, the revision process refers to instances when the seller adjusts the asking price one or more times based on the response of their listing to the market before the property is sold. The revision duration denotes the period from the first published date to the date of the last changes made to the listing price. The revision process has been studied in both classical economic research (Anglin et al. 2003; Horowitz 1992; Yavas and Yang 1995) and behavioural economic research (Bucchianeri and Minson 2013; Levitt and Syverson 2008). However, to the best of my knowledge, no empirical research has exploited information regarding the revision duration, possibly because of limitations in available data.

The price revision process can be decomposed into two distinct components: one highlighting changes in the value of the asking price, which has been extensively studied in the literature, and the other emphasising the duration that a seller spends on the revision. This seemingly straightforward dissection offers a novel approach to addressing the simultaneity problem in the price-TOM puzzle.

The listing revision duration is identified as an instrumental variable for TOM based on the following rationale. A seller may initially set a high asking price and subsequently reduce it if there is no market interest (i.e. the seller is testing the market). For instance, a seller who promptly realises that their asking price is too high and/or there is little buyer interest may have a shorter TOM than a seller who takes longer to revise the listing price downwards. The final listing price, or the last revised asking price, directly influences the TP and the remaining market duration since the last change date, while the revision duration only affects the TOM if the property quality has been controlled, which excludes any potential stigma effect of a few low-quality properties. Thus, by employing this instrumental variable, one can only account for the direction of the impact from liquidity to price and block any reverse impact of price on TOM.

However, using the price revision duration as an instrumental variable for TOM has a limitation if the model does not incorporate overpricing control. A property that is overpriced is likely to require more time or more adjustments to its asking price. As a result, without overpricing control, the coefficient of the instrumented TOM term (as an explanatory variable) in the price model is expected to be smaller than the outcome from the model specification that considers overpricing control. This implies that the estimated effect of TOM on price is likely to be negative. However, this issue does not arise when overpricing is accounted for. The empirical results in the next section confirm our expectations.

Drawing on the above discussion, the SEM that accounts for the simultaneity between TOM and price is estimated via the following 2SLS method. The first stage regressions are as follows:

$$log(T\hat{O}M) = \boldsymbol{\gamma_1 X} + \gamma_2 OPP + \gamma_3 PRD + \boldsymbol{\xi}, \tag{3.3}$$

$$log(T\hat{P}) = \boldsymbol{\lambda_1 X} + \lambda_2 OPP + \lambda_3 CTB + \boldsymbol{\varepsilon}, \tag{3.4}$$

Here, OPP denotes the overpricing proxy, PRD represents the price revision duration, and CTB is the council tax band. These equations include only the control variables and instrumental variables discussed earlier.

In the second stage, I estimate equations (3.5) and (3.6) using the estimated values of TOM from equation (3.3) and price from equation (3.4) as explanatory variables on the right-hand side, respectively.

$$log(TOM) = \beta_1 log(T\hat{P}) + \boldsymbol{\beta_2 X} + \beta_3 OPP + \beta_4 PRD + \boldsymbol{\nu}, \tag{3.5}$$

$$log(TP) = \alpha_1 log(T\hat{O}M) + \boldsymbol{\alpha_2 X} + \alpha_3 OPP + \alpha_4 CTB + \boldsymbol{\upsilon}. \tag{3.6}$$

3.5 Results

3.5.1 The Price-TOM Relationship

In TOM Model

Table 3.4 presents the results of the TOM dependent models. Columns 1 and 2 correspond to OLS estimations (using equation 3.1), whereas columns 3 to 6 show 2SLS results (using equation 3.5). Columns 1 and 3 do not incorporate the overpricing control. All models consist of fixed effects for year, month, and location. This table illustrates both the simultaneity and overpricing issues mentioned earlier. The findings reveal a positive correlation between price and TOM. Column 4 is the preferred specification for this study.

In all specifications, overpricing is found to positively affect TOM, suggesting that the higher the overpricing, the longer it takes for a property to transact. Given that overpricing negatively impacts the price, as demonstrated in both literature (Taylor 1999) and empirical results (Table 3.5), it is anticipated that the omission of overpricing control would diminish the estimated coefficient of TP in the TOM model (3.1). This is exemplified in the OLS results presented in Table 3.4. Without the overpricing proxy, as seen in column 1, the coefficient of TP (0.0508) exhibits a strong inclination towards negativity and is considerably smaller than when accounting for overpricing (0.1415) in column 2. Nevertheless, this is not apparent in the 2SLS estimations (columns 3 and 4), as the IV, CTB, utilised in the first-stage estimation (equation 3.4) is strong and exogenous.

Table 3.4: Estimations of TOM Model (Equation 3.5)

	Dependent variable: log(TOM)					
	OLS		2SLS			
	(1)	(2)	(3)	(4)	(5)	(6)
log(Transfer Price)	0.0508***	0.1415***	0.0830***	0.0906***	0.0668***	0.0668***
	(0.0027)	(0.0027)	(0.0048)	(0.0047)	(0.0039)	(0.0038)
Overpricing Proxy		1.9347***		1.9029***	1.8857***	1.8819***
		(0.0098)		(0.0101)	(0.0099)	(0.0099)
Price Revise Times	0.2573***	0.1761***	0.2576***	0.1769***	0.1773***	0.1773***
	(0.0007)	(0.0008)	(0.0007)	(0.0008)	(0.0008)	(0.0008)
Chain-free: Yes	-0.0424***	-0.0551***	-0.0407***	-0.0575***	-0.0591***	-0.0596***
	(0.0014)	(0.0013)	(0.0014)	(0.0014)	(0.0014)	(0.0014)
Price Modifier	Yes	Yes	Yes	Yes	Yes	Yes
Property Feature Variables	Yes	Yes	Yes	Yes	Yes	Yes
Energy Efficiency Variables	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
District	Yes	Yes	Yes	Yes		
Upper Tier & Unitary Authority					Yes	
Ceremonial County						Yes
IV F-test Statistic			56865.98	58963.6	68736.6	68768.86
p-value of F-test			0.0000	0.0000	0.0000	0.0000
Observations	848293	848293	848293	848293	848293	848293
Adjusted R^2	0.218	0.252	0.218	0.252	0.249	0.248
Degrees of Freedom	847946	847945	847946	847945	848158	848198
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

847946 degrees of freedom

Moreover, by accounting for overpricing, the models can minimise bias in estimating coefficients for other variables. For instance, properties that are overpriced often display a higher number of revisions to their listing prices. This can act as a confounding factor when estimating the effect of ”Price Revision Times” on TOM. By considering overpricing, there is a reduction in the estimated coefficient from 0.257 to 0.176.

Nonetheless, even with the inclusion of the overpricing control in the OLS model, the issue of bias in estimating the relationship between price and TOM remains unresolved. As demonstrated in columns 2 and 4 of Table 3.4, when comparing the outcomes of the 2SLS method (0.0906) with those of the OLS method (0.1415), the OLS method significantly overestimates the coefficient of price on TOM.

To examine the robustness of the 2SLS estimations, results of alternative 2SLS specifications are presented in the final two columns of Table 3.4, where location fixed effects are controlled with ”Upper-Tier and Unitary Authority” and ”Ceremonial County” respectively. The F-test statistics of all four 2SLS models offer compelling evidence to reject the null hypothesis that the instrumental variable used in our 2SLS models is weak. Overall, all 2SLS models consistently indicate that price has a positive impact on TOM, suggesting that sellers with higher target prices experience longer waits on the market, all else being equal.

In Price Model

Table 3.5 presents the outcomes of price dependent models. Columns 1 and 2 feature OLS estimations (equation 3.2), while columns 3 to 6 display 2SLS results (equation 3.6). Columns 1 and 3 exclude the overpricing control. All models incorporate year and month fixed effects, as well as location fixed effects. This table highlights both the simultaneity and overpricing issues previously discussed. The findings indicate that TOM is positively correlated with price. Column 4 represents the preferred specification in this study.

Table 3.5: Estimation of Transaction Price Model (Equation 3.6)

	Dependent variable: log(TP)					
	OLS		IV			
	(1)	(2)	(3)	(4)	(5)	(6)
log(TOM)	0.0082***	0.0232***	0.0247***	0.0430***	0.0445***	0.0448***
	(0.0004)	(0.0004)	(0.0018)	(0.0019)	(0.0021)	(0.0022)
Overpricing Proxy		-0.6666***		-0.7032***	-0.7018***	-0.7075***
		(0.0040)		(0.0052)	(0.0059)	(0.0061)
Price Revise Times	-0.0136***	0.0109***	-0.0178***	0.0073***	0.0070***	0.0077***
	(0.0003)	(0.0003)	(0.0005)	(0.0005)	(0.0005)	(0.0006)
Chain-free: Yes	-0.0528***	-0.0461***	-0.0520***	-0.0448***	-0.0448***	-0.0457***
	(0.0006)	(0.0005)	(0.0006)	(0.0006)	(0.0006)	(0.0007)
Price Modifier	Yes	Yes	Yes	Yes	Yes	Yes
Property Feature Variables	Yes	Yes	Yes	Yes	Yes	Yes
Energy Efficiency Variables	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
District	Yes	Yes	Yes	Yes		
Upper Tier & Unitary Authority					Yes	
Ceremonial County						Yes
IV F-test statistic			53475.20	50551.7	50777.1	50977.1
p-value of F-test			0.0000	0.0000	0.0000	0.0000
Observations	848293	848293	848293	848293	848293	848293
Adjusted R^2	0.854	0.858	0.853	0.858	0.817	0.800
Degrees of Freedom	847946	847945	847946	847945	848158	848198

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

In Table 3.5, a negative correlation between overpricing and price is observed. Specifically, the more severe the overpricing, the lower the price. Omitting overpricing in the model has a significant impact on the estimation of both TOM and "Price Revision Times." As can be seen in both the OLS results (column 1) and the 2SLS results (column 3), the estimated coefficients of TOM are much smaller when overpricing is not controlled for, compared to models where overpricing is accounted for (columns 2 and 4). Similarly, it is found that the coefficient of "Price Revision Times" is negative when overpricing is not taken into account, but becomes positive when overpricing is controlled for. This suggests that sellers who have revised their asking prices more frequently achieve higher prices. However, it is important to note that sellers who overprice tend to modify their asking prices more often, which can confound the effect of "Price Revision Times." When comparing the 2SLS results in column 4 to the OLS estimation in column 2, it is evident that the OLS estimation tends to underestimate the coefficient of TOM in the price model.

To assess the robustness of the price dependent model estimation, results of two alternative 2SLS models are presented in the final two columns of Table 3.5, wherein location fixed effects are controlled with "Upper-Tier and Unitary Authority" and "Ceremonial County" respectively. The F-test statistics of all four 2SLS models provide strong evidence to reject the null hypothesis that "Price Revision Duration" is a weak IV. These results further support the finding that TOM has a positive effect on price, signifying that staying longer on the market leads to a higher TP, all else being equal. This could be related to an exposure effect that a longer TOM increases the seller's probability of encountering a higher offer from buyers.

More Robustness Tests

To further scrutinise the robustness of the aforementioned findings, I conducted a bootstrap robustness test. Additional analyses were performed on three subsets obtained through random sampling with replacement from the full sample. The sample sizes for each subset were 300,000, 500,000, and 800,000, respectively. Then, I re-estimated both preferred TOM and price models (as outlined in equations 3.5 and

3.6) using the three subsets. Subsequently, I repeated this process 500 times for each model and recorded the coefficient of TOM in the price model and the coefficient of price in the TOM model, along with their p-values. It is worth noting that these subsets were re-sampled from the full dataset in each of the 500 iterations.

Table 3.6: Robustness Tests

	TOM Model		Price Model	
	log(TP)	p-value	log(TOM)	p-value
Re-sampling size: 300,000				
Min.	0.0618	0.00e+00	0.0315	0.00e+00
1st Qu.	0.0841	0.00e+00	0.0407	0.00e+00
Median	0.0901	0.00e+00	0.0432	0.00e+00
Mean	0.0903	5.35e-18	0.0430	2.38e-26
3rd Qu.	0.0961	0.00e+00	0.0451	0.00e+00
Max.	0.1213	2.66e-15	0.0542	1.19e-23
Re-sampling size: 500,000				
Min.	0.0711	0.00e+00	0.0340	0.00e+00
1st Qu.	0.0866	0.00e+00	0.0413	0.00e+00
Median	0.0907	0.00e+00	0.0429	0.00e+00
Mean	0.0906	9.05e-34	0.0429	3.40e-47
3rd Qu.	0.0945	0.00e+00	0.0447	0.00e+00
Max.	0.1088	4.46e-31	0.0507	1.70e-44
Re-sampling size: 800,000				
Min.	0.0777	0.00e+00	0.0360	0.00e+00
1st Qu.	0.0875	0.00e+00	0.0415	0.00e+00
Median	0.0906	0.00e+00	0.0429	0.00e+00
Mean	0.0906	6.27e-61	0.0429	6.67e-83
3rd Qu.	0.0938	0.00e+00	0.0443	0.00e+00
Max.	0.1058	2.95e-58	0.0496	1.83e-80

The summary of these results is presented in Table 3.6. These findings demonstrate the robustness of the preferred 2SLS models in this study and their ability to produce consistent estimates despite the potential presence of sample selection bias.

3.5.2 Chain-free Sellers and Agency Costs

The analysis now turns to examining how the financial characteristics of sellers relate to the relationship between price and TOM. In regards to sellers' chain status, an intriguing finding has emerged when combining the empirical results in both the liquidity and price models. Intuitively, when other factors are equal, a chain-free property sells faster if it is sold at the same price compared to in-chain homes (Table 3.4). However, Table 3.5 reveals that it is sold at a 4-5% lower price for the same TOM. The modelling results in this section indicate that a chain-free property has a 4-5% lower initial asking price, which potentially suggests the presence of agency cost issues.

A property chain refers to the sequence of transactions that must take place for a property to be sold. It typically involves multiple buyers and sellers, each of whom is dependent on the sale or purchase of a property further up or down the chain. For example, if an individual is selling a property and wants to purchase another property, they are dependent on the sale of their current property to proceed with the new one's purchase. Likewise, the person they are selling the property to may be dependent on the sale of their current property to purchase the one they are interested in, and so on. This creates a chain of buyers and sellers, each of whom is dependent on the sale or purchase of another property to move forward with their own transaction. The longer the chain, the more complex and time-consuming the process can be.

Conversely, the term "chain-free" in an advertisement indicates that the seller is not dependent on the sale of their current property to proceed with their next steps, which can include buying another property or not. This can make the process of buying or selling a property faster and less complicated as it eliminates the need to coordinate the sale of multiple properties simultaneously. Properties described as "chain-free" are often more appealing to buyers as they allow for a more flexible and efficient buying process. Thus, if a property is advertised as "chain-free", it can be viewed as an indicator that the seller is not financially constrained and that the buying process is likely to be more streamlined.

There are primarily four categories of sellers who might classify their properties as "chain-free." The first category comprises homeowners with an alternative residence who have no plans to purchase a new property, due to reasons such as emigration, altered personal circumstances, or the sale of a deceased relative's estate. The second category involves financial institutions disposing of properties acquired through repossession, probate, or equity release. The third category is constituted by home builders, who are typically represented by companies rather than individual sellers, resulting in a chain-free status. Lastly, the fourth category includes professional investors who procure properties as part of a larger portfolio rather than as a primary residence.

There are several potential explanations for why chain-free sellers tend to sell their properties at 4-5% less for the same marketing duration. One explanation is that these sellers may have a heightened sense of urgency to sell their properties, leading them to accept a lower sale price to facilitate a quicker sale. Another explanation is that chain-free sellers may have lower bargaining intensity, potentially due to being less financially constrained or influenced by other behavioural factors. A third explanation could be that estate agents may advise chain-free sellers to set a lower initial listing price to increase the likelihood of a sale, as noted by Levitt and Syverson (2008)[6], potentially indicating agency cost issues. In contrast, for sellers in a property chain, the process of selling is interconnected among multiple sellers. The speed at which the chain progresses is often determined by the slowest seller's selling process. This means that even if an agent wants to move quickly, the entire chain's progress is limited by the seller who is taking the longest time to complete their sale. As a result, an "in-chain" seller is less likely to suffer from the agency cost issue.

The first theory suggests that chain-free sellers accept lower sale prices in exchange for expedited sales. However, it is improbable that all chain-free sellers in the market share an identical preference for a swifter sale. The data employed in this study reveals that approximately 32% of transactions were labelled as "chain-free." Although it is conceivable that some sellers, such as those emigrating or selling on behalf of a deceased relative, might favour a rapid sale with a reduced TP, these instances are likely to constitute a minority and have an insignificant influence on the modelling

[6] This research pointed out that current standard home sale contracts create a potential conflict of interest between sellers and their agents. While agents only earn a small percentage of the final sale price, they shoulder significant upfront costs (showings, open houses, marketing). This can incentives agents to prioritise a quick sale, even if it means accepting a lower offer for the seller.

outcomes. The findings of this study indicate that this behaviour is a general tendency in the market. Moreover, even assuming all chain-free sellers are driven by a preference for faster sales, the results presented in Table 3.5 imply that, on average, when sellers have not secured a quicker sale, they would have achieved a lower sale price.

Glower et al. (1998) developed a housing search model positing that home sellers eager to sell promptly will establish a lower listing price and accept earlier, reduced offers. However, the empirical test in their research fails to support the claim that sellers influence the marketing of their properties through strategic setting of the initial listing price. Consequently, they propose that motivated sellers progressively decrease the asking price to accelerate the sales process. They conclude that sellers eager for a swift sale only impact TP, but not the listing price markup, indicating that they do not establish a lower initial listing price. This suggests that sellers targeting rapid sales would set an initial listing price indistinguishable from that of other sellers. Based on these rationale, it can be deduced that the urgency of sale is not a valid explanation if a negative 'chain-free' effect on the initial listing price exists.

The second hypothesis posits that the lower sale price observed among chain-free sellers arises from their low bargaining intensity. However, this hypothesis can be refuted if being "chain-free" has a negative impact on the initial listing price. Rationally, sellers not driven by the necessity for a swift sale would lack any incentive to set a lower listing price on the first day, prior to negotiating with prospective buyers. At a market level, these sellers would not wish to decrease their potential sale price by establishing a lower listing price. Consequently, if there is a negative effect of being "chain-free" on the initial listing price, then the reduced bargaining intensity of chain-free sellers cannot be regarded as a valid explanation.

Hence, determining whether being listed as "chain-free" has a negative impact on the initial listing price is crucial for identifying the plausible explanation among the three assumptions mentioned above.

The test results, displayed in Table 3.7, demonstrate that being listed as chain-free does indeed exert a negative influence on the initial listing price. Specifically,

Table 3.7: Effect of Chain-free on Initial Listing Price

	Dependent variable:					
	log(Initial Listing Price)					
	(1)	(2)	(3)	(4)	(5)	(6)
Chainfree: Yes	−.0451***	−.0467***	−.0450***	−.0446***	−.0443***	−.0413***
	(.0006)	(.0006)	(.0006)	(.0006)	(.0006)	(.0006)
log(TOM)	.0216***					
	(.0005)					
Overpricing Proxy	.1567***	.1939***				
	(.0041)	(.0040)				
Price Revise Times	.0108***	.0115***	.0189***			
	(.0004)	(.0004)	(.0004)			
Price Modifier: Guide Price	.0583***	.0582***	.0589***	.0554***		
	(.0040)	(.0040)	(.0040)	(.0040)		
Price Modifier: Offers Around	.0619***	.0622***	.0637***	.0616***		
	(.0041)	(.0041)	(.0041)	(.0041)		
Price Modifier: Offers Over	.0364***	.0368***	.0361***	.0359***		
	(.0041)	(.0041)	(.0041)	(.0041)		
Price Modifier: Price on Request	.1098***	.1064***	.1087***	.1094***		
	(.0085)	(.0085)	(.0085)	(.0085)		
Price Revise Duration	.0001***	.0001***	.0001***	.0003***	.0003***	
	(.00001)	(.00001)	(.00001)	(.000004)	(.000004)	
Property Feature Variables	Yes	Yes	Yes	Yes	Yes	Yes
Energy Efficiency Variables	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Location Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	848,293	848,293	848,293	848,293	848,293	848,293
Adjusted R^2	.8540	.8536	.8532	.8529	.8527	.8516

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.8: The Mediation of Initial Listing Price between Chain-free and Transaction Price

	log(TP)	log(TP)
Chain-free:Yes	-0.0462***	-0.0167***
	(0.0033)	(0.0015)
log(Initial Listing Price)		0.8899***
		(0.0039)
Num.Obs.	848293	848293
R^2	0.919	0.985
Property Characteristics	Yes	Yes
Energy Efficiency Variables	Yes	Yes
FE: CTB	Yes	Yes
FE: Location	Yes	Yes
FE: Time	Yes	Yes
Note:	*p<0.1; **p<0.05; ***p<0.01	

the findings reveal that a chain-free property has a 4-5% lower initial listing price, which is the same magnitude as the loss in TP for chain-free sellers. This effect remains robust across models that control for different sets of variables related to the seller's characteristics (Price Revise Times, Price Modifier, and Price Revise Duration). These findings imply that the agency cost assumption is the most plausible explanation. To furnish further evidence, I examine the causal mechanism of this assumption, where the initial listing price acts as a mediator of the negative relationship between chain-free status and TP.

Table 3.8 illustrates the function of the initial listing price as a mediator between chain-free status and TP, employing hedonic model regressions. The negative effect of being chain-free on the TP diminishes considerably from 0.0462 to 0.0167 when the initial listing price is incorporated into the model. This lends support to the hypothesis that chain-free status impacts the TP by initially affecting the listing price.

Drawing on the above analysis, this study asserts that agency costs constitute the primary factor contributing to the lower initial listing price for properties not part of a chain, resulting in reduced sale prices over the same marketing period. In contrast, "in-chain" sellers lack the incentive or means to sell swiftly at a lower price, as they face greater financial constraints and the sale's pace largely depends on the progress of other sales in the chain. As a result, they are less likely to be influenced by agents to set a lower initial listing price, being more risk-averse than chain-free sellers. Consequently, they tend to establish a higher initial listing price and encounter fewer agency-related issues compared to chain-free sellers.

Similarly, from the viewpoint of information asymmetry between sellers and estate agents, Buchak et al. (2020) discovered that iBuyers—intermediaries who buy and sell residential real estate through online platforms to accelerate the transaction process—provide liquidity to households by allowing them to circumvent a protracted sale process and generate a 5% spread. This suggests that the information asymmetry is valued at a comparable amount (4-5% of the price) in this study, which indirectly supports the agency costs explanation. However, due to the limitation of open-source data, private information about individual buyers and sellers is unobserved in the data. Future research endeavours can consider incorporating the characteristics of both sellers and buyers, providing a more nuanced understanding that may contribute to a deeper exploration of the counter intuitive findings associated with 'chain-free' transactions.

3.6 Conclusion

This chapter examines the long-standing puzzle of the relationship between price and TOM in the empirical literature. Search theory predicts a positive correlation between TOM and price. However, before 2016, empirical findings concerning this relationship were notably inconsistent, with over 400 estimations producing positive, negative, or insignificant results. Despite three studies published between 2016 and 2019 specifically addressing this issue, the outcomes remain divergent.

To tackle this puzzle, I compiled a multi-source, integrated dataset using open data and information obtained from a task-tailored web crawler on publicly accessible websites, as detailed in chapter 2. The final sample comprises roughly one-third of all full market value residential transactions sold and registered with HM Land Registry in England between January 2018 and March 2021. This dataset is unique in its scope and allows for an accurate and robust analysis of the England housing market.

Given the simultaneous nature of the relationship between price and TOM, relying exclusively on OLS estimations would lead to biased results. To account for this endogenous relationship, I modelled the relationship with a SEM and applied a 2SLS estimation procedure. The efficacy of 2SLS in controlling for simultaneity relies on the quality of the instrumental variables employed. It is widely acknowledged that 2SLS is biased towards the OLS estimator when instruments are weakly correlated with the endogenous regressor. This bias is evident in the negative or insignificant slope coefficients on TOM in the price dependent model reported in a previous study (Hayunga and Pace 2019).

Traditionally, the weak instrument test relies on the F-statistics of the first-stage auxiliary regression. However, this test proves largely uninformative in terms of both size and bias, as emphasised in recent literature (Young 2022). Thus, alongside the weak instrument F-test, I also examine the theoretical relevance of the IVs and the endogenous variables, which allows us to evaluate the quality of the IVs and the validity of the results.

I identify two novel instrumental variables for the SEM, which have not been employed in previous studies. I use the CTB as the IV for the price, as it is assessed by the VOA using scientifically-based criteria at a county level, rendering it exogenous to all participants involved in the transaction. For TOM, I utilise the ”Price Revision Duration” as its IV, as it demonstrates a positive correlation with TOM but does not directly affect the TP. This variable effectively blocks any reverse impact from TP to TOM in the estimation of the price-dependent model.

Another challenge in modelling the price-TOM relationship is the endogeneity issue arising from the omission of an overpricing measure, as discussed in a previous theoretical study (Taylor 1999). Despite this, prior research on the relationship between price and TOM has not yet identified a direct solution to this problem. To address this issue, I devised a proxy for overpricing, which also accounts for the negotiation friction between buyers and sellers. The modelling results reveal that overpricing has a positive effect on TOM and a negative effect on the price, in line with theoretical predictions. Furthermore, my findings confirm that the omission of overpricing control leads to significant confounding issues when estimating the price-TOM relationship.

My findings demonstrate that the relationship between price and TOM is positive and simultaneously determined, in accordance with search theory. Additionally, I observe that chain-free sellers—those not required to sell their current property before purchasing a new one—tend to set lower initial asking prices and agree to lower transaction prices, all else being equal. I suggest that agency costs are the primary factor contributing to the lower initial listing price for properties not part of a chain. In contrast, "in-chain" sellers, who face greater financial constraints and have their sale speed dependent on the progress of other sales in the chain, lack the same incentive or means to sell quickly at a lower price. As a result, they are less likely to be persuaded by agents to set a lower initial listing price, being more risk-averse. Consequently, they tend to have a higher initial listing price and experience fewer agency-related issues compared to chain-free sellers.

Chapter 4

Transaction Taxes and Housing Market Dynamics

4.1 Introduction

The property transaction tax, also referred to as real estate transfer tax or stamp duty, is a levy imposed on the sale of real estate. This tax is typically paid by the buyer at the time of sale. In England, residential properties are subject to the Stamp Duty Land Tax (SDLT), with tax rates varying based on factors such as whether it is a first-time or second home purchase, the price of the property, and other criteria. As of 2022, the tax rate ranges from zero to 12 percent, with the tax applied to most properties above a house price threshold of £125,000. Consequently, this tax directly affects the majority of home-buyers in England.

Governments may implement this tax for various reasons, with the primary motivation being to generate revenue for public programmes and services such as education, healthcare, and infrastructure development. The UK has had a stamp duty in place since the 16th century, and it remains a significant source of revenue for the government. Over the years, there has been a considerable increase in the number of property transfers subject to stamp duty. In 1997, 49 percent of all transactions were subject to stamp duty, compared to 75 percent in 2015. In fact, net residential receipts in the UK have nearly tripled over the last decade, from £2.95 billion in

2008-09 to £8.42 billion in 2019-20[1].

In addition to generating revenue for the government, a property transaction tax can also serve as a means of regulating the real estate market. By making it more challenging for speculators to engage in excessive buying and selling of property, a property transaction tax can help prevent the formation of housing bubbles. Moreover, a property transaction tax can encourage long-term ownership of property. By imposing a tax on the sale of property, the government can incentivise individuals and families to retain their properties for extended periods, fostering more stable and sustainable communities and ensuring that property owners contribute to the public good, rather than using their property solely for personal gain.

As with any tax on transactions, property transaction taxes can negatively impact market activity and lead to sub-optimal market behaviour. High transaction taxes might discourage mutually beneficial trades and reduce the liquidity of the housing markets, making it more difficult for individuals to buy and sell property, which can hinder economic growth and development. In situations where the property market is already struggling, a property transaction tax can act as a further deterrent to buying and selling, potentially exacerbating the problem.

Past research has criticised transaction taxes as being inefficient because they can negatively affect mobility, preventing people from moving, and resulting in adverse effects on employment and productivity (Hilber and Lyytikäinen 2017; Van Ommeren and Van Leuvensteijn 2005). A property transaction tax can create a disincentive for individuals to move to new areas or upgrade to larger homes, negatively impacting the economy by limiting opportunities for individuals to relocate for work or take advantage of new housing developments. It can also make it difficult for individuals to move to areas with better schools or upgrade to homes that better suit their needs.

Furthermore, it is important to consider the institutional context of each country when evaluating the impact of transaction taxes. The frequency of property transactions can vary greatly across regions and types of households. However, there is no

[1] Data from: `https://www.gov.uk/government/statistics/quarterly-stamp-duty-land-tax-sdlt-statistics`

strong economic justification for imposing excessive transaction taxes on frequently traded residential properties (Adam 2011).

Consequently, the decision to implement a property transaction tax is multifaceted, necessitating an in-depth examination of potential costs and benefits. Although it can provide a valuable revenue source for the government and assist in regulating the real estate market, it is crucial to implement it in a way that does not overly burden individuals and families. To minimise adverse effects on the economy and lower-income individuals, the property transaction tax should be meticulously designed and implemented in collaboration with stakeholders from the real estate industry and the wider community.

In December 2014, the UK government replaced the widely criticised "slab" tax structure with a new "slice" system, akin to income tax, in which only the portion of a property price falling within certain bands is subject to tax. However, compared to the old system, there is limited knowledge about the effects of the new SDLT. This thesis aims to address this gap by thoroughly analysing the effects of the revised transaction tax system on the UK housing market through a quasi-natural experiment based on the 2020 stamp duty holiday (hereafter referred to as SDH) in the UK.

The COVID-19 pandemic's outbreak and the imposition of the first nationwide lockdown, from March to May 2020, led to a standstill in the UK's property market. As part of a package of job creation measures, a temporary tax holiday was introduced with immediate effect on 8th July 2020. The initial announcement stated that the SDH would be in effect until 31st March 2021. However, shortly before this deadline for the SDH, the SDH was extended for three more months and gradually phased out.[2].

The SDH involved suspending the payment of transaction tax on properties valued at less than £500,000. The policy increased the nil rate band of stamp duty for residential transactions from £125,000 to £500,000. Consequently, home movers and first-time buyers did not have to pay transaction tax if their property was valued up to

[2] Subsequently, England, Wales, and Northern Ireland decided to extend the SDH until 30th June 2021. After 30th June 2021, SDLT was not paid for properties below £250,000 until the end of September 2021. A return to its original rates was announced for the 1st of October 2021

£500,000. If their property was worth more, they only paid the tax on the proportion of the value exceeding this threshold, based on rates in a tiered tax system. Investors, such as buy-to-let buyers, also benefited from the SDH policy and paid 3% more on the portion of the TP above £500,000.

The 2020 SDH was not the first instance of a tax holiday being introduced in the UK. In 2008, to counter the adverse effects on the housing market caused by the Global Financial Crisis, an SDH was introduced. Several studies (Besley et al. 2014; Best and Kleven 2017; Hilber and Lyytikäinen 2017) have examined transaction taxation on the UK housing market under the old slab system. These studies have demonstrated that a sudden and unexpected removal of transaction taxes can result in a short-term increase in sales when housing supply is inflexible. Besley et al. (2014) demonstrates that this effect is short-lived and is offset once the tax is reintroduced, leading to the conclusion that market participants time transactions. Best and Kleven (2017) show that the suspension of a 1% stamp duty tax rate in a certain price range boosts market activity by 20%; this is followed by a reversal of about 8% one year after the SDH during 2008-09. While research has been conducted on the 2008 SDH, the method for calculating the property transaction tax has changed since the introduction of a new slice tax system in December 2014. Under the new tax regime, one would expect that the effects of an SDH on housing market activity and prices may differ from those of the previous slab system's.

This thesis addresses the research gap by thoroughly examining the newly improved stamp duty system, using the extensive datasets constructed in Chapter 2. The objective is to quantify the impact of the SDH policy under a tiered structure on housing prices, transaction and listing volumes, liquidity, as well as to determine the distribution of the tax burden between buyers and sellers. At the time of writing this thesis, this is the first study to (i) comprehensively explore the effects of the slice stamp duty in the UK, and (ii) assess the effect of the recent SDH on housing market activity from the above perspectives.

In this chapter, I first investigate the changes in price and surplus for both buyers and sellers. Following previous research (Besley et al. 2014; Kopczuk and Munroe

2015), I propose a Nash bargaining model for the new tiered tax system. The model predicts higher transaction prices during the SDH. It utilises the bargaining power between the seller and the buyer to explain the tax incidence between these two actors. A seller will have more surplus if the property is traded during the SDH, while for a buyer, the surplus change primarily depends on the magnitude of the changes in bargaining power and tax savings. This also helps to interpret some of the empirical results in this chapter.

Empirically, I find that the SDH significantly influences short-term housing market activity, increasing monthly listings and transactions by 60% and 53%, respectively. Moreover, both asking and transaction prices experience a surge due to the SDH. The average increase in transaction prices ranges from 1.9-2.5%. This increase approximately corresponds to the tax savings realised by home movers replacing their main residence. Consequently, the savings from the tax break are entirely distributed to the sellers. For first-time buyers, the price increase resulting from the SDH exceeds twice the tax-saving. This suggests that the entire amount of tax savings from the SDH is passed on to the sellers in the form of elevated prices. This observation aligns with some previous studies (Dachis et al. 2012; Davidoff and Leigh 2013).

One of the intended outcomes of the SDH policy, introduced in the wake of the Covid outbreak, was to boost economic activity by increasing expenditure on housing-related goods and services. Nevertheless, these findings reveal that the SDH did not free up extra funds for home movers to spend, even though they are the primary agents who could have increased consumption of housing-related goods and services. Policymakers might have relied on earlier studies (Besley et al. 2014) analysing the 2008 SDH under a slab tax system, which found that 40% of the tax savings were distributed to sellers. The 2020 SDH did not benefit home movers and first-time buyers.

Furthermore, this thesis reveals the presence of strategic market timing behaviour of agents in response to the SDH policy. Sellers rapidly list their properties to capitalise on the SDH, considering the policy's limited eight-month duration. Additionally, based on data indicating that a transaction takes an average of six months[3], this chapter demonstrates that the number of monthly new listings declines four months prior to the SDH deadline. This implies that potential sellers are deterred from entering the market, as the probability of finding a buyer and completing a deal within less than four months is considerably reduced.

The strategic market timing behaviour is also evidenced by the evolution of price spread, measured as the difference between the final asking price and the TP. The closer the transaction is to the deadline of the SDH, the closer is the TP to the asking price. Controlling for the final listing price, this suggests a shift in the bargaining positions of buyers and sellers, with buyers feeling more pressure to meet sellers' asking prices. Therefore, sellers' bargaining position strengthens as the deadline of the SDH approaches and homes sell above asking prices.

Given that the SDH was introduced during the peak of the Covid pandemic, this thesis also evaluates the extent to which the SDH was linked to potential relocation effects due to remote working. The findings suggest that market participants utilised the SDH to move away from large metropolitan areas. This conclusion is reinforced by the observation that the SDH led to a shift in demand from flats to houses (see also Petkova and Weichenrieder 2017 and Fritzsche and Vandrei 2019). The majority of flats in England are situated in densely populated urban areas and lack outdoor space, rendering them less attractive purchases when remote working is a viable option.

4.2 The Institutional Setup

4.2.1 The Stamp Duty Land Tax in England

The stamp duty has a long history in the UK tax system. Introduced in 1694 to finance the war against France, it was initially applied to vellum, parchment, and paper transactions. The tax levied on stamp duty was easy to identify and measure, while few other potential taxes were straightforward to implement (Adam 2011). By

[3] This is the time between the first date a property is listed and the recorded date on the transaction deed.

1808, housing had been added to the list of items subject to stamp duty.

The SDLT is unappealing, as any charge on transactions reduces expected benefits by discouraging mutually advantageous trades, ensuring that properties are not held by those who value them most. By law, buyers must pay the full stamp duty at the time of purchase and cannot fund it through mortgages. This increases the amount required for a down payment on a property, discouraging people from moving and potentially contributing to labour market inflexibility. It also encourages people to reside (and businesses to operate) in properties of a size and location they might not have chosen otherwise. Moreover, the frequency with which a house is traded varies greatly, but there is no compelling economic justification for taxing more frequently traded housing multiple times (Adam 2011).

Prior to December 2014, the stamp duty in the UK had a slab structure, which was reinstated in the first Budget following the Labour Party's election in 1997. Buyers were required to pay a tax rate based on the total purchase price, resulting in higher tax rates for a specific transaction being applied to the full price, not the portion above the relevant threshold. This led to the discouragement of property values slightly above a threshold, creating substantial incentives for buyers and sellers to agree on prices just below the relevant thresholds. Best and Kleven (2017) demonstrate that the discontinuity in SDLT rates in the UK generates bunching around the notch in the sale distribution just below the levels that trigger a higher rate on the entire price, as well as a gap immediately above them. Similarly, Slemrod et al. (2017) provide evidence of this notch and bunching effect by examining a series of transaction tax revisions in Washington DC that introduced discontinuous jumps in tax obligations. Furthermore, Kopczuk and Munroe (2015) exploit the discontinuity in tax burden caused by the so-called mansion tax imposed in the states of New York and New Jersey. They argue that the bunching effect can extend up to 10% of the threshold value in the sale price distribution.

The current SDLT features a slice structure or tiered structure, introduced by the Conservative-Liberal Democrat coalition government in December 2014. The new stamp duty rates apply only to specific bands of the property price, resembling the progressive structure of income tax. For instance, based on the tax rates before the SDH, as shown in Table 4.1, a property transferred at £300,000 would incur a tax of £5,000 (0% for the first £125,000, 2% for the next £125,000, and 5% for the remaining £50,000). The Treasury asserted that the new tiered structure would reduce stamp duty for 98% of payers[4]. The tiered system also eliminates the artificial discontinuities in property price distribution by reducing the cliff-edge jumps at the boundaries of rate ranges (Scanlon et al. 2017).

A surcharge tax was introduced in April 2016 for property buyers in England and Wales. It added a 3% surcharge on each stamp duty band for buy-to-let[5] properties and second homes. This aims to favour owner-occupiers over investors or second-home buyers. According to the Treasury, the higher rates of SDLT on new residential property acquisitions form part of the government's commitment to supporting home-ownership and first-time buyers[6].

Over time, the number of property transfers subject to stamp duty has increased, with the percentage of transactions subject to stamp duty rising from 49% in 1997 to 75% in 2015. This growth has made the SDLT a significant revenue source for the UK government, with net residential receipts tripling from £2.95 billion in 2008-09 to £8.42 billion in 2019-20.[7].

4.2.2 The 2020 Stamp Duty Holiday

In 2020, the COVID-19 pandemic caused a sharp slowdown in the UK's property market and economy[8]. In April 2020, there were 38,060 property transactions[9], 46,230

[4] See, https://www.gov.uk/government/publications/stamp-duty-reforms-factsheet

[5] Buy-to-let properties can be identified by a buy-to-let mortgage, which is granted to small landlords, i.e., homeowners who want to buy property to let it out.

[6] See, https://www.gov.uk/government/consultations/consultation-on-higher-rates-of-stamp-duty-land-tax-sdlt-on-purchases-of-additional-residential-properties/higher-rates-of-stamp-duty-land-tax-sdlt-on-purchases-of-additional-residential-properties

[7] Data from: https://www.gov.uk/government/statistics/quarterly-stamp-duty-land-tax-sdlt-statistics

[8] From the 23rd of March to the 13th of May, the UK enforced its first national lockdown, asking individuals to "stay at home." Most estate agencies had to stop viewings and operations altogether.

[9] See, Coronavirus: UK property sales hit record low in April: https://www.bbc.co.uk/news/business-52752475

in May[10], which was less than half of the number recorded for the same month the previous year. Furthermore, in June of that year, house prices fell by 0.1% compared to June of the previous year, marking the first annual drop since December 2012.[11]

On the 8th of July 2020, the Chancellor of the Exchequer announced a temporary increase in the nil rate band for residential housing sales as part of a job-creation package during a statement to the House of Commons on the state of the economy amid the COVID-19 pandemic. The objective of this SDH was to stimulate housing market activity and boost the economy by driving demand for housing-related goods and services. According to the Treasury press release[12], it would achieve this by: (i) immediately lowering purchase costs; (ii) limiting the SDH period to incentivise buyers to shift their plans forward to enjoy the benefits; and (iii) freeing up money for buyers to spend on housing-related goods and services.

Table 4.1: Stamp Duty Land Tax Rates before or during the Stamp Duty Holiday

Transfer value	SDLT rate	
	Before SDH	During SDH
Up to £125,000	0	0
£125,001 to £250,000	2%	0
£250,001 to £500,000	5%	0
£500,001 to £925,000	5%	5%
£925,001 to £1.5 million	10%	10%
More than £1.5 million	12%	12%

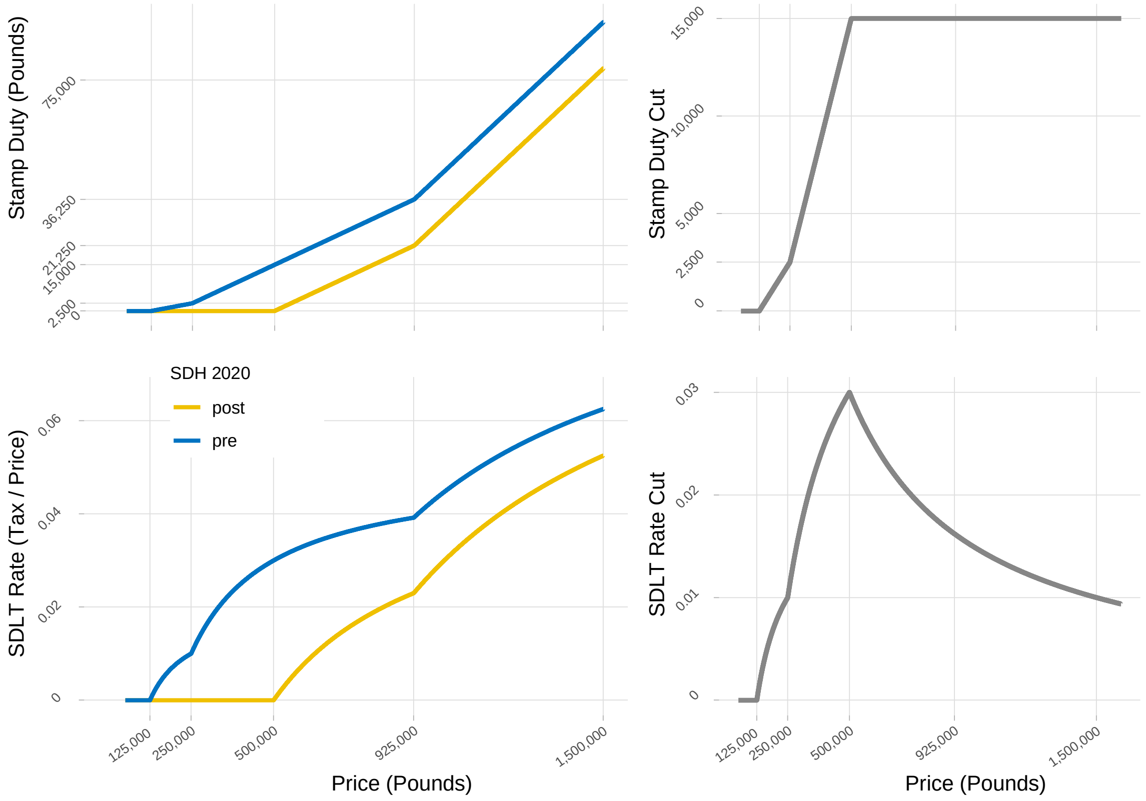
Note: The nil rate band for first-time buyers stands at £300,000. For second-home buyers, a 3% addition is made to each band rate, equivalent to adding 3% of the total price to the stamp duty. Home movers are exempt from the 3% additional rate if they are replacing their primary residence.

[10] See, Coronavirus: House sales plummeted by 50% in May: https://www.bbc.co.uk/news/business-53148678

[11] See, Coronavirus may have huge impact on property markets: https://www.bbc.co.uk/news/business-52977890

[12] https://www.gov.uk/government/speeches/a-plan-for-jobs-speech; https://www.gov.uk/government/news/stamp-duty-holiday-continues-to-help-hundreds-of-thousands-of-jobs-after-further-213-boost-in-september

Figure 4.1: Stamp Duty Tax Reduction



Note: The horizontal axes of the four graphs represent the TP. The two graphs on the left depict the stamp duty and SDLT rates before (blue line) and after (yellow line) the introduction of the 2020 SDH. The panels on the right show the corresponding tax cut (top-right), provided in British Pounds, and the tax rate cut, measured in percentages of the price (bottom-right).X

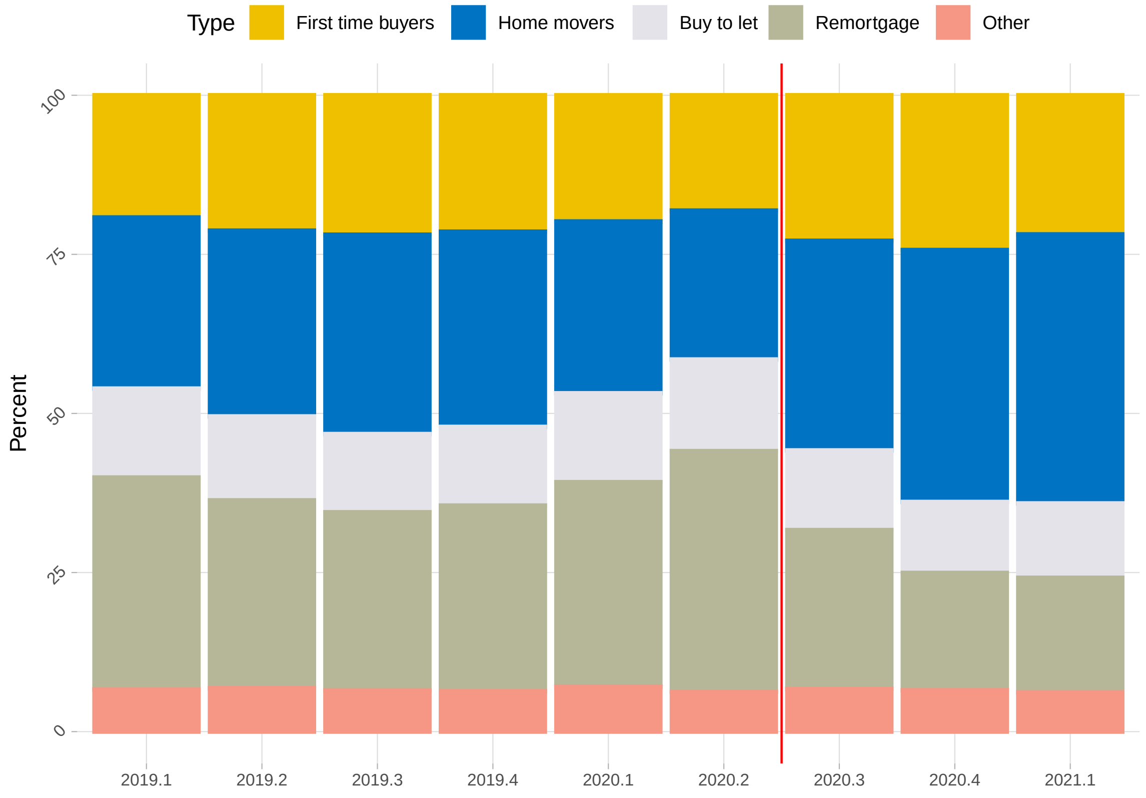
The "During SDH" column in Table 4.1 displays the updated tax rates during the SDH. The initial £125,000 nil rate band of SDLT was increased to £500,000. Figure 4.1 illustrates the effective SDLT rate and the reduction in rate based on the total TP. Properties valued over £500,000 are limited to tax savings of £15,000, as seen in the top-right corner of the figure. Properties priced at £500,000 receive the largest tax rate saving of 3%, as depicted in the bottom-right corner of Figure 4.1. The stamp duty and the stamp duty rate are continuous functions of property prices, and the new tiered tax structure does not have notches, which were previously used as an identification strategy in UK market studies. A different identification technique will be used in this thesis, as described in Section 4.5.

During the SDH, the majority of transactions were made by first-time buyers and home movers. Figure 4.2 provides a breakdown of the purpose of mortgage loans. From the onset of the SDH, a decrease is observed in both the buy-to-let and remortgage segments, while an increase is seen in both first-time buyers and home movers. In the fourth quarter of 2020, the share of remortgages was 18.5%, representing a decrease of 10.7 percentage points compared to the fourth quarter of 2019 and the lowest level since 2007. The share of buy-to-let purposes was 11.2%, a decrease of 1.2 percentage points from the fourth quarter of 2019. The share of first-time buyers was 24.3%, an increase of 2.9 percentage points compared to the fourth quarter of 2019. Meanwhile, the share of home movers rose to 39.6%, an increase of 8.9 percentage points compared to the previous year, marking the highest share for home movers since the third quarter of 2010[13].

The property market in England experienced an annual price increase of 8.5% in December 2020, elevating the average property value to £269,150. Regional data indicates that the North West exhibited the highest annual price increase, surging by 11.2%, whereas London had the lowest growth at 3.5%[14]. The tax holiday was initially scheduled to conclude on March 31, 2021, but was subsequently extended to June 30, 2021 (in England, Northern Ireland, and Wales), followed by a nil rate band

[13] Data source: Financial Conduct Authority, https://www.fca.org.uk/data/commentary-mortgage-lending-statistics-q4-2020

[14] https://www.gov.uk/government/news/uk-house-price-index-for-december-2020



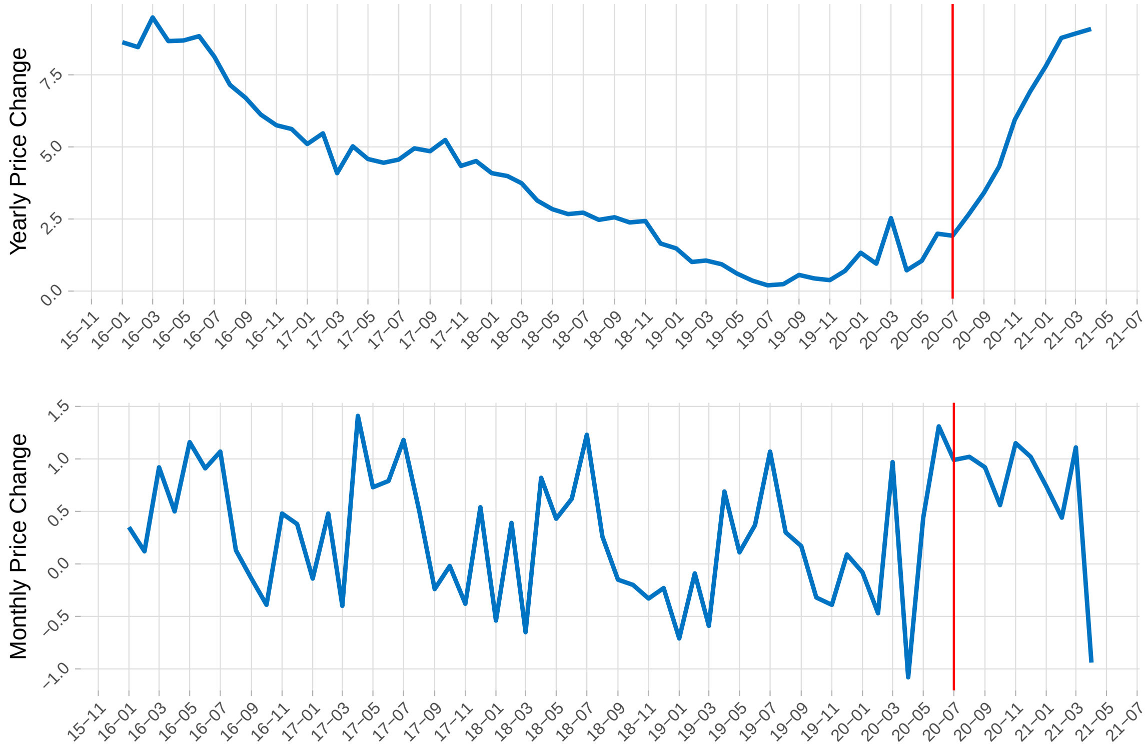
source: Financial Conduct Authority and Bank of England

Figure 4.2: Breakdown of Mortgage by Purpose

Note: The figure illustrates the distribution of mortgage loan purposes before and during the stamp duty holiday. Prior to the SDH (before the red vertical line), less than half of the loans were designated for first-time buyers and home movers. With the commencement of the SDH, an immediate decline in the share of remortgages and buy-to-let purposes can be observed, highlighting the shift in the mortgage market during the stamp duty holiday.

of £250,000 until September, before returning to pre-holiday tax rates on October 1st.

Figure 4.3 illustrates the average property price fluctuations in England from January 2016 to April 2021. The upper plot depicts the annual changes, demonstrating a 4-year downward trend prior to the SDH, accompanied by a brief resurgence from January to March 2020. Nevertheless, the first lockdown in March 2020 impeded this progress, causing prices to stabilise at 2% in June 2020. During the SDH, the annual price changes dramatically escalated to 8.9% by March 2021. In the lower panel, property prices experienced a more significant increase between April and July 2020 compared to other months. Typically, the monthly change peaks between April and July, declines sharply in August-September, and then oscillates around zero during winter. However, this seasonal pattern was disrupted, and prices continued to rise steadily throughout the SDH.



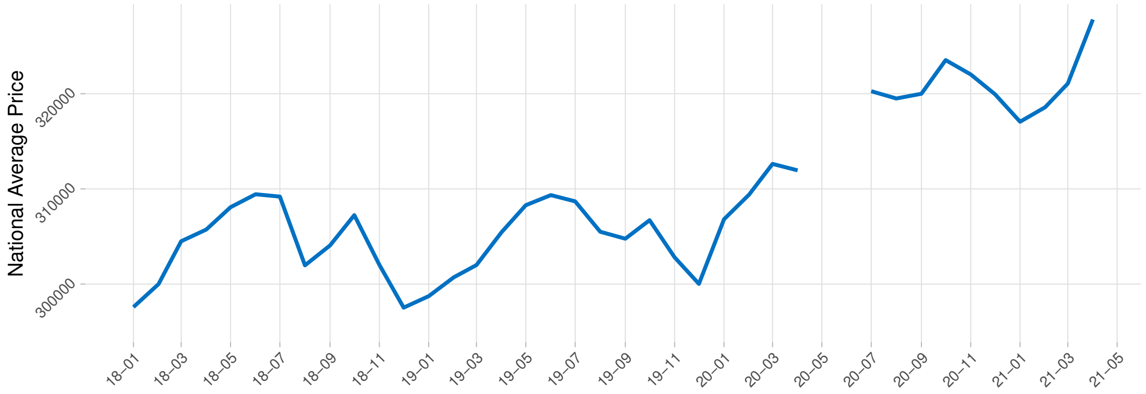
Data source: UK House Price Index

Figure 4.3: Price Changes (Percentage) in England

Note: The upper plot in the figure presents the average property price in England from January 2016 to April 2021, whereas the lower plot illustrates the variations in the average property price throughout the same time frame. The vertical line marks the onset of the SDH. The first national lockdown due to Covid-19 occurred between late March and June in 2020.

Figure 4.4 presents the Rightmove[15] Asking Price Index from January 2018 to April 2021. Throughout the SDH period, a notable surge in asking price can be observed. The typical seasonal pattern, characterised by peaks during May-July and declines from July to December, is absent during the SDH. Moreover, the number of listings also escalates to a higher level compared to the same period in previous years.

[15] Rightmove is the UK's largest online real estate portal and property website. Data for May and June 2020 was suspended by Rightmove due to the Covid lockdown.



from Rightmove Asking Price Index, data in May and June 2020 was suspended by Rightmove due to pandemic and national lockdown.

Figure 4.4: Asking Price in England

Note: This figure illustrates the Asking Price Index from Rightmove spanning January 2018 to April 2021. Notably, the data points for May and June 2020 were suspended by Rightmove owing to the COVID-19 pandemic.

Regarding identification in this thesis, it is crucial that the policy is not anticipated by market participants and materialises as an exogenous shock. We can assume that the market had no foreknowledge of the 2020 SDH. Figure 4.5 depicts the Google Trends data for the search volume of ”stamp duty holiday” between January 2019 and March 2021. As demonstrated in the figure, there was minimal or no search volume for the keyword prior to July 2020, signifying that the market held little expectation of the SDH.

Only a few days prior to the SDH announcement, media speculations suggested that the nil rate band of SDLT could increase to as much as £500,000[16], while other commentators voiced concerns that lowering stamp duty tax rates would significantly and negatively affect the stock market.[17]

Before July 2020, the primary discussion regarding the SDLT pertained to the 2% surcharge proposed for non-UK resident buyers in the 2020 Spring Budget and the 2019-21 Finance Bill, which was presented in early March 2020. However, the Budget

[16] See, ”Stamp duty ’holiday’ to help rebuild economy”, Times, 6 July 2020`https://www.thetimes.co.uk/article/stamp-duty-holiday-to-help-rebuild-economy-2t0rhgphg`; See, ”UK chancellor to help the young in summer statement”, Financial Times, 6 July 2020`https://www.ft.com/content/502f31d6-c3f1-4608-a892-f1e5a35d6d33`

[17] See, ”Stamp duty plan risks bringing sales to a standstill for months”, Times, 7 July 2020; ”Sunak risks losing his sparkle with ill-timed stamp duty holiday”, Times, 7 July 2020

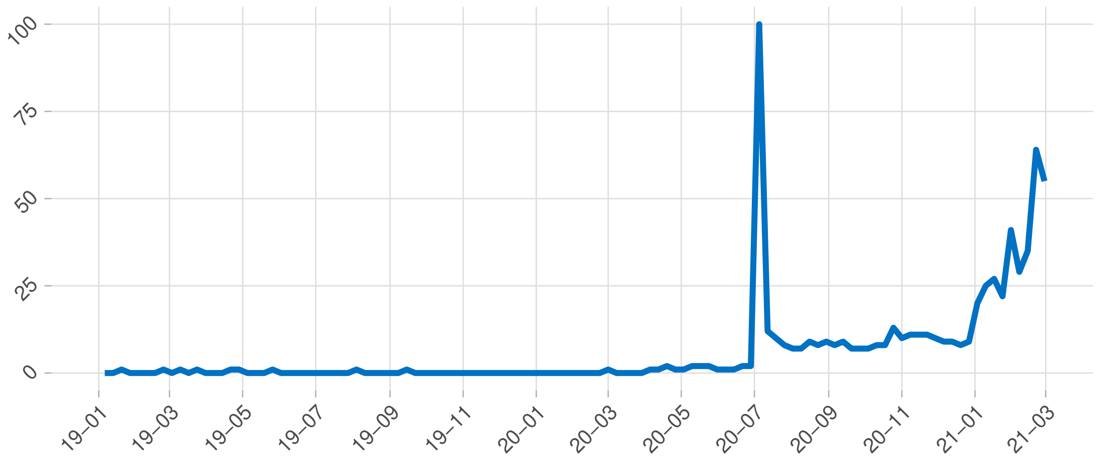


Figure 4.5: Google Trends for ’stamp duty holiday’
Note: The plot exhibits the search volume index (scaled between 0 and 100) of ”stamp duty holiday” on Google Search from January 2019 to March 2021. It reveals that there was minimal or no search volume for the keyword before July 2020, implying that the market possessed no prior awareness or anticipation of the SDH.

did not introduce any further changes to the existing SDLT regime. In late March 2020, the implementation of lockdown restrictions bore potential tax implications for homeowners who concurrently owned two properties due to relocation. Owing to the slowdown caused by the Covid pandemic, households might have been unable to sell their previous main residence within the three-year window necessary to qualify for a refund of the 3% tax surcharge for second homes. To address this issue, the government introduced new clauses in June 2020, allowing for an extension of the 3-year time limit under specific circumstances.

4.3 Economies of Property Transaction Taxes

The literature on transaction taxation predominantly focuses on two aspects. Firstly, the impact of taxation on market activity and household mobility. In most cases, the tax is levied on buyers and affects market demand, subsequently influencing transactions and household mobility. Secondly, the effect of tax on property prices and the economic incidence of the tax. Tax incidence refers to the distribution of the tax burden between the buyer and seller (or between employer and employee, or between firms and consumers). The total amount that buyers are willing to pay, including price and stamp duty, should not be affected by the imposition of transaction taxes. Consequently, the price must fall, and the tax burden partially falls on the sellers(Adam 2011).

4.3.1 Transaction Volume and Mobility

The literature generally concurs that imposing taxes on property transactions results in a decrease in sales, while the temporary removal of these taxes leads to an increase in market turnover.

Several studies (Besley et al. 2014; Best and Kleven 2017; Hilber and Lyytikäinen 2017) have examined transaction taxation on the UK housing market under the old, slab system. The 2020 SDH is not the first time a tax holiday has been introduced in the UK. To counteract the adverse effects on housing markets resulting from the Global Financial Crisis, a SDH was introduced in 2008. Best and Kleven (2017) demonstrate that suspending a 1% stamp duty tax rate within a specific price range enhances market activity by 20%; this is followed by a reversal of approximately 8% one year after SDH, suggesting that market participants re-time their transactions. Since the aggregate housing stock cannot respond to tax policy changes in the short run, they interpret this stimulus effect as increasing sales of existing housing stock.

According to Besley et al. (2014), it led to an 8% increase in the number of sales in the relevant pricing window (influenced by SDH), but it was offset by a significant drop that followed when the SDH ended. This implies that the effect on volumes is a short-term re-timing of transactions. In March 2021, the government extended the SDH until 30th June. This was followed by a nil rate band of £250,000 per property until the end of September 2021. From 1st October 2021, the tax rates returned to the pre-SDH ones. Examining residential property transaction data in Figure 4.6, we can observe a spike in June and October 2021 and declines in August, September, and November 2021. This thesis does not investigate whether the 2020 SDH is associated with re-timing of transactions that would have occurred anyway or provided an additional boost given the insufficient sample for the post-SDH period in my data.

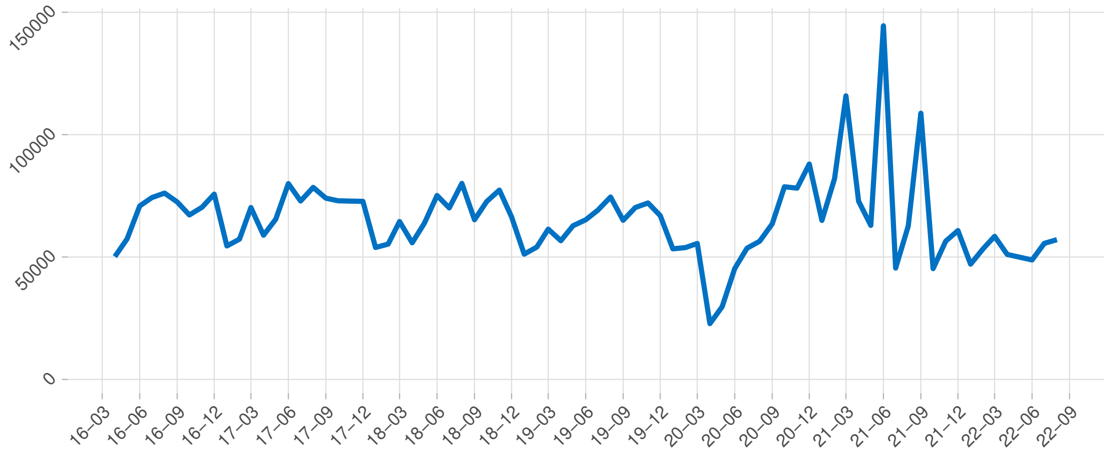


Figure 4.6: Residential Property Transaction Volume in England by Month

Petkova and Weichenrieder (2017) and Fritzsche and Vandrei (2019) examine the German market to estimate the effect of property transaction taxes on sales volumes. Petkova and Weichenrieder (2017) use annual indices of property transactions to study the effect for single-family homes and apartments separately. They find that the tax significantly affects the number of transactions only for single-family homes. According to Fritzsche and Vandrei (2019), a 1 percentage point increase in the real estate transaction tax reduces single-family house transactions by approximately 7%.

Davidoff and Leigh (2013) analyse the effects of transaction tax with a sample of 25,111 observations from Australia. They highlight a significant empirical issue in evaluating the effects of taxes on the market in the absence of a quasi-experimental context. Stamp duties are endogenous concerning the purchase price of a home. Therefore, they develop an IV and use a 2SLS method to quantify the effect. In their preferred specification, a 10% increase in stamp duty lowers housing turnover by 3%.

Dachis et al. (2012) utilise an unanticipated introduction of land transfer tax in Toronto to estimate the tax's effect on transactions using a regression discontinuity model. Comparing the number of sales across the boundary of Toronto, they find a 1.1% increase in tax causes a 15% decrease in transactions. Using the above econometric strategy requires the sales just outside the boundary of tax change (the control group) to be unaffected by the tax policy introduced in a certain location (the treatment group), but this could be violated if housing sorting occurs from the affected area to the unaffected area. Slemrod et al. (2017) find no evidence of a timing effect in the volume of house sales when studying the notched tax rate changes in Washington DC.

Another strain of research assesses the effect of transaction taxes on household mobility. Hilber and Lyytikäinen (2017) find that the tax only has a negative impact on short-distance (10 km or less) moves using UK survey data of around 20,000 households. Using data for the entire Finnish population from 2005 to 2016, Eerola et al. (2021) find negative mobility effects in both short-distance (less than 50 km) and long-distance (more than 50 km) re-locations. Han and Sheedy (2022) show that transaction taxes negatively affect owner-occupiers' mobility and distort housing tenure choices. The tax falls more heavily on owner-occupiers as they are expected to transact more frequently and are then expected to pay a new transaction tax every time a new property is purchased. This makes existing owner-occupiers more tolerant of poor match quality and decreases the moving rate.

4.3.2 Price and SDLT Incidence

The economic incidence of a tax may differ from its statutory incidence. Economic incidence refers to the individual or group of individuals who ultimately bear the actual cost of the tax, while statutory incidence pertains to those responsible for physically remitting a specific tax to the government. It is natural to assume that stamp duty does not adversely affect the seller, as it is typically paid by the purchaser (as is the case in many countries, including the UK). However, this may not necessarily be true and can depend on the bargaining positions of both the buyer and seller. Consider a fixed short-term supply of houses[18]. The house price will then be influenced by demand factors, i.e., the amount purchasers are willing to pay. Imposing a transaction tax should not impact the total amount (price plus stamp duty) buyers are willing to

[18] This is a plausible assumption in the housing market because obtaining planning permission and completing construction can take years.

pay, all else being equal. Consequently, in order to maintain a constant total cost for the buyer, the price should decrease. As a result, the tax burden, at least partially, rests on the sellers.

Best and Kleven (2017), Kopczuk and Munroe (2015), and Slemrod et al. (2017) examine behaviour around a price notch to study the actual tax incidence between buyers and sellers. Davidoff and Leigh (2013) analyse this using exogenous variation in stamp duty rates in Australia, while Besley et al. (2014) investigate it using the 2008 SDH in the UK as a quasi-experiment. Their findings indicate that the economic incidence of the tax on sellers ranges from 40% to over 100%. According to Besley et al. (2014), buyers received 60% of the surplus generated from the SDH, implying that 40% of tax incidence falls on sellers. Best and Kleven (2017) and Slemrod et al. (2017) reveal that the quantity of bunching below the notch roughly equals the "missing" volume for properties priced just above the notch, suggesting that buyers and sellers share the real incidence of tax equally. Davidoff and Leigh (2013), in their preferred specification, find that a 10% increase in stamp duty reduces property prices by 4-5%, with prices falling by the full amount of the tax. This implies that the economic incidence of the transaction tax falls entirely (100%) on the seller, consistent with the findings of Dachis et al. (2012). However, Kopczuk and Munroe (2015) discover that the volume of "missing" transactions above the price notch is significantly greater than the volume of bunching below the notch, indicating that sellers bear more than 100% of the real incidence of the tax in their scenario, which cannot be explained by tax evasion.

In summary, the evidence suggests that property transaction taxes negatively impact transaction volumes and prices. Nonetheless, little is known about the effect of the tax on other aspects of the market, such as TOM, asking prices, and bid-ask price spreads. This thesis contributes to the existing literature by shedding light on housing market dynamics more broadly, employing a quasi-natural experiment and a rich proprietary dataset for the UK.

4.3.3 A Proposed Model for Bargaining Power

To further understand the effects of the 2020 SDH on market behaviour, I modify the Nash bargaining model employed in previous studies (Besley et al. 2014; Kopczuk and Munroe 2015). In prior research, this model was designed to accommodate a slab-structured tax system. Under this system, buyers are subject to a tax rate that is applied to the full price of the property, with the rate being determined by the price bracket in which the transaction falls. This results in a lump-sum tax, represented by an additional rate multiplied by the full price, being imposed only on buyers whose property price exceeds a certain threshold, which in turn leads to notches in the price distribution. I have adapted the aforementioned model to the new slice tax system under the 2020 SDLT regime.

Consider the case where a buyer and a seller bargain over the price. The buyer's valuation of the house is b with a bargaining power of α, and the seller's valuation is s with a bargaining power of $1-\alpha$; This can be described as a single match (b, s) with the outcome determined through Nash bargaining. The negotiated TP between the buyer and the seller is denoted as p. Trade takes place only if $b > p \geq s$.

In Besley et al. (2014), the model maximises the function $(b-\tau p-p)^{\alpha}(p-s)^{1-\alpha}$ with respect to p, where τ represents a change in the tax rate for properties in the treatment group and is set to zero for the control group. Similarly, in Kopczuk and Munroe (2015), the model maximises the function $(b-T-p)^{\alpha}(p-s)^{1-\alpha}$ with respect to p, where T represents the aforementioned lump-sum tax. These two models were utilised to compare similar properties that fall below or above a threshold in the slab-structured tax system.

In this model, given the sale price, the buyer ends up with surplus $S^B = b + K - (1+\tau)p$,[19], and the seller ends up with surplus $S^S = p - s$, where τ is the marginal stamp duty rate and K is a constant determined by the price band in which p falls into. For example, if the TP is in the band $[£250,001, £500,000]$,

[19] The constant K is defined as $K := \tau\lfloor p \rfloor - sdlt(\lfloor p \rfloor)$ where $\lfloor p \rfloor$ means the left bound of the price band that p falls into subtract 1. For example, if $p \in [250,001, 500,000]$, then $\lfloor p \rfloor = 250,000$. $sdlt(\lfloor p \rfloor)$ represents the amount of the stamp duty if the transaction price is $\lfloor p \rfloor$.

then the buyer will pay 0% for the first £125,000 of the total price, 2% for the second £125,000, and 5% for the price above £250,001. So the buyer's surplus is $S^B = b - 2500 - \tau(p - 250000) - p = b + K - (1 + \tau)p$, where $K = 10000$ and $\tau = 0.05$. Similarly, the trade takes place only if both S^B and S^S are positive. We assume the Nash bargaining with buyer's weight α and seller's weight is $1 - \alpha$. The price maximisation equation is given as:

$$\underset{p}{\text{argmax}}(b + K - (1 + \tau)p)^{\alpha}(p - s)^{1-\alpha} \tag{4.1}$$

where τ and K are fixed numbers when the price falls within a certain band. Then it yields the following formula for the price (see Appendix 4.7 for the proof):

$$p = \frac{1 - \alpha}{1 + \tau}(b + K) + \alpha s \tag{4.2}$$

Correspondingly, the seller's and the buyer's surplus is expressed as:

$$S^S = (1 - \alpha)(\frac{b + K}{1 + \tau} - s) \tag{4.3}$$

$$S^B = \alpha(b + K - (1 + \tau)s) \tag{4.4}$$

The following table shows the corresponding τ and K for each price band both before the SDH and during the SDH. I denote τ_0 and K_0 for the period before SDH and τ_1 and K_1 for the time during SDH. In the case of $\tau = 0$, it means there is no stamp duty as the transaction price falls into certain price bands.

If the SDH policy does not affect the buyer's and seller's valuations and bargaining power (I will relax some of these assumptions later), I define $p_1 = \frac{1-\alpha}{1+\tau_1}(b + K_1) + \alpha s$ and $p_0 = \frac{1-\alpha}{1+\tau_0}(b + K_0) + \alpha s$ as the price during and before the SDH respectively. For property values over £500,000, $\tau_0 = \tau_1$ while $K_0 < K_1$, thus $p_1 > p_0$ always holds. The change in price as a result of the SDH for properties valued between £125,001 and £500,000, which was the target range of the SDH policy, is given as:

$$\Delta p = p_1 - p_0 = \frac{1 - \alpha}{1 + \tau_0}(b\tau_0 - K_0) > 0. \tag{4.5}$$

Since a deal takes place only if a buyer's valuation b is larger than the agreed price

Table 4.2: Parameters in Formula (4.19)

Price Band	Before SDH		During SDH	
	τ_0	K_0	τ_1	K_1
Up to £125,000	0	0	0	0
£125,001 to £250,000	2%	£2,500	0	0
£250,001 to £500,000	5%	£10,000	0	0
£500,001 to £925,000	5%	£10,000	5%	£25,000
£925,001 to £1.5 million	10%	£56,250	10%	£71,250
More than £1.5 million	12%	£86,250	12%	£101,250

Note: This table shows the parameters τ and K for each price band in formula (4.19). Noticed that the maximum difference between K_0 and K_1 is the maximum saving one can have from the SDH policy, which is £15,000.

p, $\Delta p > 0$ holds as $b\tau_0 - K_0 > p\tau_0 - K_0 \geq 0$ for any $p \in [125001, 500000]$. Therefore, under the above assumptions, the model shows the TP will be higher if the trade happened during the SDH.

I now relax the above assumptions to more realistic ones, it is plausible to assume a fixed short-term supply of houses. Therefore house prices will mainly be driven by demand factors. What follows from this is the first assumption that the buyer's bargaining power should decrease during the SDH because it boosts the demand and leads to a seller's market, denoted by $\alpha_1 < \alpha_0$ (therefore, seller's power increases, denoted as $1 - \alpha_1 > 1 - \alpha_0$). The second assumption is that the seller's valuation does not change during the SDH, which is the key assumption to Besley et al. (2014). The third assumption claims that the buyer's valuation does not decrease during the holiday, denoted by $b_1 \geq b_0$.

Under these assumptions, let $p_1^* = \frac{1-\alpha_1}{1+\tau_1}(b_1 + K_1) + \alpha_1 s$ denote the agreed price during the SDH, for a price in the range $[125001, 500000]$, the change in the TP is

$$\begin{aligned} p_1^* - p_1 &= (1-\alpha_1)b_1 - (1-\alpha_0)b_0 - (\alpha_0 - \alpha_1)s \\ &\geq (1-\alpha_1)b_0 - (1-\alpha_0)b_0 - (\alpha_0 - \alpha_1)s \\ &= (\alpha_0 - \alpha_1)(b_0 - s) > 0. \end{aligned}$$

Accordingly, the change of the price due to the SDH is:

$$\Delta p^* = p_1^* - p_0 > p_1 - p_0 > 0. \tag{4.6}$$

For a property valued over £500,000, the change in the TP is:

$$\Delta p^* = p_1^* - p_0 = \frac{1-\alpha_1}{1+\tau_1}(b_1 + K_1) + \alpha_1 s - \frac{1-\alpha_0}{1+\tau_0}(b_0 + K_0) - \alpha_0 s. \tag{4.7}$$

Noticing that $K_1 = K_0 + 15000$ (see Table 4.2) and $b_1 \geq b_0$ (this means that buyers' valuation of a property during the SDH is no less than their valuation before the tax holiday), then we have:

$$\begin{aligned}\Delta p^* &\geq \frac{1}{1+\tau_0}[(1-\alpha_1)(b_0 + K_0 + 15000) - (1-\alpha_0)(b_0 + K_0) - (\alpha_0 - \alpha_1)s(1+\tau_0)] \\ &= \frac{1}{1+\tau_0}[(1-\alpha_1)15000 + (\alpha_0 - \alpha_1)(b_0 + K_0 - s(1+\tau_0))].\end{aligned} \tag{4.8}$$

Therefore we have:

$$\Delta p^* \geq \frac{1}{1+\tau_0}[(1-\alpha_1)15000 + (\alpha_0 - \alpha_1)(b_0 + K_0 - p_0(1+\tau_0))]. \tag{4.9}$$

If the transaction happened before SDH, we can claim the buyer's surplus is positive, denoted as $S^B = b_0 + K_0 - p_0(1+\tau_0) > 0$. Consequently it means the property will be traded with a higher price during the SDH because $\Delta p^* = p_1^* - p_0 > 0$ holds under the relaxed assumptions.

As demonstrated by equation (4.3), the parameters presented in Table 4.2, and the assumptions outlined above, it is straightforward to confirm that the seller's surplus will be greater when the property is traded during the SDH (see Appendix 4.7 for the proof). This is reasonable as a portion of the tax-saving from SDH goes to the seller. On the other hand, for the buyer (as represented by equation 4.4), their surplus may vary depending on the changes in bargaining power and the tax savings. In other words, the buyer's surplus is dependent on the proportion of tax-saving allocated to the buyer and the increase in price during SDH.

In conclusion, the proposed Nash bargaining model suggests that the price of a property traded during the SDH will be higher than if it were sold before the SDH.

4.4 Data

This thesis focuses solely on residential property transactions in England rather than the entire UK market for several reasons. Firstly, among the four nations of the UK, England's housing market transactions account for more than 90% of the UK's SDLT receipts since 2000[20]. This number has increased to over 97% in recent years, as the SDLT was fully devolved to Scotland and Wales in 2015 and 2018. Consequently, total UK SDLT receipts do not include receipts from these two nations. Secondly, the devolution of SDLT has led to differences in property transaction tax holidays between England and the other UK nations. Following the SDH in England, the Scottish Government temporarily raised the nil threshold from £145,000 to £250,000 on 15th July 2020, ending the holiday on 31st March 2021 without extension. The Welsh Government temporarily increased the nil-rate band from £180,000 to £250,000 on 25th July 2020, with an extension until 30th June 2021. Lastly, the listings information was collected only for England, and the other key databases solely contain properties in England and Wales.

This chapter evaluates the effects of the SDH on (i) property listing and transaction volumes, and (ii) sales details, such as TP, initial listing price, TOM, and price spread. The former quantifies the number of transactions or listings at a national level for each month from January 2018 to March 2021 by aggregating PPD or listing data, respectively. This leads to a small sample size ranging between 78 and 234 observations.

For the purpose of investigating the impact of SDH on sales details, several property-level datasets containing records from April 2018 to March 2021 are used

[20] For more information on this data, see `https://www.gov.uk/government/collections/stamp-duties-statistics`

in this chapter. To analyse the initial listing price, two samples have been employed. The first, the Listings-EPC-CTB dataset (referred to as the large sample in listing price analysis), comprises 1.35 million observations and includes only matched records from listings data, EPC data, and CTB data. The second, called the baseline sample, consists of linked records from all data sources but has fewer observations, totalling 675,701. However, this sample benefits from having additional variables available from PPD compared to the large Sample.

To analyse the TP, another large sample (PPD-EPC-CTB), consisting of 2.14 million transactions, is used. This sample is obtained through matching data from PPD, EPC, and CTB and represents over 92% of the population transactions recorded with the Land Registry. Additionally, another baseline sample has been employed, linking records from all data sources with a smaller sample size of 814,937, but offering a broader range of variables. This sample is also used for analysing the TOM and price spread. This trade-off arises due to the limited availability of variables during the matching process when using different combinations of all data sources. Moreover, comparing the modelling results from the baseline sample and large sample can provide insight into the robustness of the models and whether the smaller baseline sample is representative of the population.

It is important to note that the observations contained in the two datasets used for analysing the asking price are determined by the initial listing date, while observations in the two datasets for analysing TP are determined by the transfer date. Consequently, these are two distinct baseline samples, which are two different subsets of the full Listings-PPD-EPC-CTB dataset constructed in Chapter 2. In particular, the baseline sample for listing price analysis is also a subset of the baseline sample for sales details analysis, because some of the transferred properties in the sample period were listed before April 2018.

In addition, properties are classified into three rural-urban categories in above datasets, using the 2011 Rural-Urban Classification (RUC) for output areas (OAs)[21]

[21] OAs are the smallest geographic unit for which Census data are available. Their geographical size will vary depending on the population density. OAs were built from clusters of adjacent unit postcodes. For the 2011 Census, England was divided into 171,372 OAs which, on average, have a

in England[22]. The classification is shown in Figure 4.7[23]. In 2011 in England, 82.4% of the population resided in urban areas. The urban areas are divided into urban conurbation (39% of the population) and urban city and town (43.4% of the population). In Figure 4.7, dark grey areas represent the urban conurbations, which consist of a number of metropolises, cities, large towns, and other urban areas that have merged through population growth and physical expansion to form a continuous urban or industrially developed area, such as Greater London, Greater Manchester, and the West Midlands conurbation. Meanwhile, the light grey areas represent urban cities and towns. These urban conurbations often exhibit a polycentric urbanised structure, with transportation systems that have developed to create a single urban labour market or travel-to-work area. It is worth noting that while rural areas occupy 85% of the land area, only 18% of properties are located in rural areas.

Table 4.3 presents the summary statistics of the baseline sample used for analysing the TP, TOM, and price spread, encompassing 814,937 observations. Numerical variables are characterised by their minimum, 25th percentile, mean, median, 75th percentile, and maximum values, while categorical variables are described by count and frequency percentages. Table 2.2 in Chapter 2 provides definitions for these variables. The average TP stands at £311,607, with a median of £259,900, indicating a slight right skew. Prices range from a minimum of £15,000 to a maximum of £2,050,000, with the 25th and 75th percentiles at £172,500 and £385,000, respectively. The average initial listing price is £327,692, with a distribution resembling that of transaction prices. However, the maximum initial asking price is notably higher at £3,750,000. Price spread distribution reveals considerable fluctuations across transactions, with an average price spread between the final listing price and the TP of £9,348. The spread's minimum is -£700,000, with the 25th percentile at 0, the median at £5,000, the 75th percentile at £12,500, and the maximum at £1,950,000. On average, the final listing price of £320,955 is nearly 2% lower than the initial listing price, displaying

resident population of 309 people.

[22] https://www.gov.uk/government/statistics/2011-rural-urban-classification

[23] The source of the figure https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/propertysalesinruralandurbanareasofenglandandwales/september2011toyearendingseptember2015.

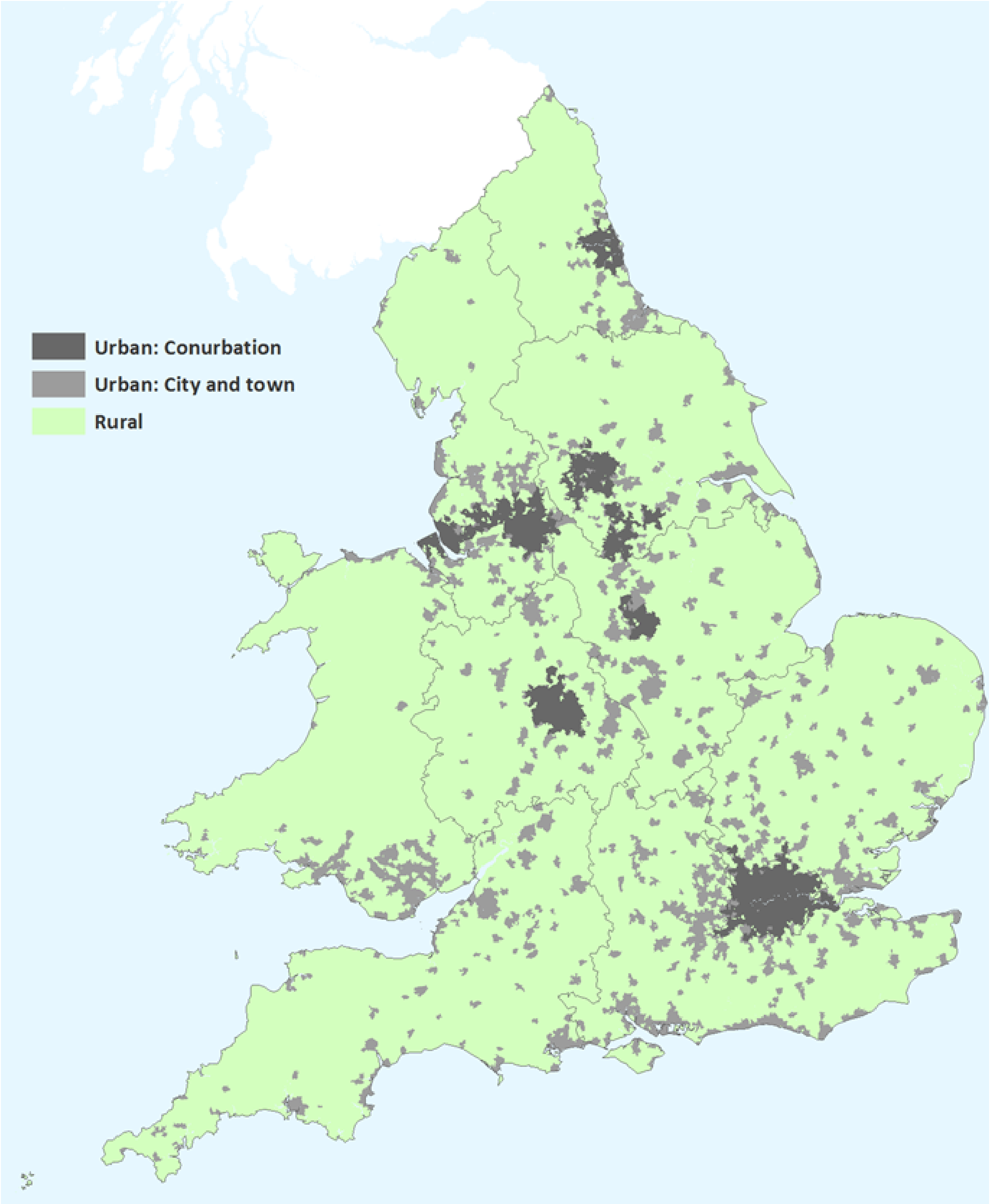


Figure 4.7: Rural-Urban Classification, Source: Office for National Statistics

a similar distribution. The average TOM is 222 days, or approximately 7.4 months, with a median of 171 days, signifying a right skew; some properties take significantly longer to transact. The 75th percentile is 265 days, or 8.8 months, while the 25th percentile is 119 days, or around 4 months. This indicates that, to benefit from the tax savings within the SDH period, sellers needed to list their properties as soon as they became aware of the SDH, and buyers had to promptly agree on deals. However, it is important to note that, within the English housing system, agreed prices are not legally binding; either the seller or buyer can withdraw from the transaction or renegotiate the price at any time, rendering the transaction process highly uncertain until the final completion date.

Table 4.3: Summary Statistics of Continuous Variables

Statistic	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
TP	15,000	172,500	311,606.90	259,900	385,000	2,050,000
TOM	1	119	221.72	171	265	1,163
Price spread	−700,000	0	9,347.76	5,000	12,500	1,950,000
Initial listing price	12,000	180,000	327,691.50	270,000	400,000	3,750,000
Final listing price	12,000	175,000	320,954.70	265,000	399,000	3,500,000
Total floor area	26	73	98.78	89	114	446
Num. habitable rooms	1	4	4.88	5	6	11
Num. open fireplaces	0	0	0.15	0	0	40
Current energy efficiency	1	57	62.92	64	70	142
Potential energy efficiency	1	78	80.71	82	85	142
Environment impact current	1	51	59.08	60	68	136
Environment impact potential	1	73	77.95	80	84	139
Energy consumption current	−257	193	255.75	241	302	1,831
Energy consumption potential	−338	88	127.56	114	151	1,417

Observations: 814,937

The average total floor area is 99 square metres, and properties typically feature 5 habitable rooms. As shown in Table 4.4, approximately 91.6% of transactions involve houses, while a mere 8.4% are for apartments, which aligns with the aforementioned property dimensions. A scant 0.14% of transactions pertain to newly-built properties. Garages are included in 47.5% of transactions, driveways in 41.9%, and gardens, which are present in almost every house, account for 88.3%.

The energy efficiency ratings range from 1 to 142 (Table 4.3), with higher ratings indicating better efficiency. Current Energy Efficiency is concentrated between 57 and 70 (mean 62.92), indicating a balanced distribution. Potential Energy Efficiency, with a range of 78 to 85 (mean 80.71), suggests a positive outlook. The gap between current and potential ratings (57-70 vs. 78-85) highlights an opportunity for improvement, emphasising the need for strategic interventions and sustainability practices. The environmental impact ratings range from 1 to 136 for current environmental impact and 1 to 139 for potential environmental impact, with higher ratings indicating a lower environmental impact. Current Environmental Impact is concentrated between 51 and 68 (mean 59.08), suggesting a moderate environmental impact. Potential Environmental Impact, with a range of 73 to 84 (mean 77.95), indicates a potential reduction in environmental impact. Comparing the current and potential environmental impact ratings (51-68 vs. 73-84) reveals a scope for improvement, highlighting the potential for a more sustainable and environmentally friendly operation. Similarly, many properties demonstrate considerable potential for reducing energy consumption through enhanced energy efficiency measures.

Table 4.4 reveals that 50% of properties are situated in cities and towns, 28.5% in urban conurbation areas such as large metropolitan zones like London, and 20% in rural locations.

Table 4.5 compares the means of key variables of interest, including transaction prices, initial listing prices, price spreads, and TOM, for various data splits. Comprehensive summary statistics can be found in Tables 5.1, 5.2, 5.4, and 5.3 in the Appendix. When divided by RUC, urban conurbation areas exhibit the highest transaction and asking prices, as well as the largest price spreads, while urban city and town areas display the smallest values. On average, houses are more expensive than flats and are associated with smaller price spreads. Moreover, both transaction and asking prices are higher, on average, during the SDH compared to the period before; the price spread is narrower during the SDH. These findings will be corroborated later by regression models.

Table 4.4: Count and Frequency of Categorical Variables

Statistics	Count	Frequency
RUC		
... Rural	162,911	20%
... Urban City&Town	419478	51.5%
... Urban Conurbation	232548	28.5%
Property type		
... House	746,298	91.6%
... Flat	68,639	8.4%
Freehold	703,402	86.3%
New built	1,136	0.14%
Price modifier		
... Fixed price	2833	0.35%
... Guide price	599690	73.6%
... Offers around	60542	7.4%
... Offers over	150983	18.5%
... Price on request	889	0.11%
Chainfree	262,872	32.3%
Garage	387,335	47.5%
Driveway	341,591	41.9%
Garden	719,763	88.3%
CTB		
... A	104907	12.9%
... B	156904	19.3%
... C	201268	24.7%
... D	163358	20%
... E	106592	13.1%
... F	52595	6.5%
... G	27925	3.4%
... H	1388	0.2%

4.5 Methodology

I evaluate the effects of the SDH using a DiD design that enables comparison of the evolution of outcomes between two groups (treated and control) while accounting

Table 4.5: Comparing the Means of the Interested Outcome Variables across Various Data Splits

By RUC			
	Rural	Urban City&Town	Urban Conurbation
	Mean	Mean	Mean
Transaction price	320,692	282,277	358,150
Initial listing price	339,202	296,025	376,749
Price spread	10,951	7,803	11,011
TOM	241	216	218
By Property Type			
		House	Flat
		Mean	Mean
Transaction price		315,507	269,200
Initial listing price		331,292	288,550
Price spread		9,161	11,383
TOM		218	257
Transacted before vs during SDH			
		Before	During
		Mean	Mean
Transaction price		297335	344261
Initial listing price		313769	359544
Price spread		9478	9051
TOM		204	262
Listed Before vs During SDH			
		Before	During
		Mean	Mean
Transaction price		305207	361799
Initial listing price		322230	370520
Price spread		9879	5518
TOM		232	138

for individual characteristics, fixed differences across groups and locations, and all other time-fixed changes. The treatment group comprises properties subject to the SDH, i.e. those priced above £125,000, while the control group consists of properties unaffected by the tax holiday, i.e. those priced at £125,000 or below. The DiD compares transactions within each group before and during the implementation of the SDH in July 2020.

The identification strategy relies on the parallel trends assumption, which posits that trends in volume and price should not correlate with the SDLT reduction in the absence of the SDH. Consequently, one can calculate the pre-treatment difference between the treated and control groups, as well as the post-treatment difference; then, under the parallel trends assumption, the difference between these two calculated differences represents the causal effect of the treatment, which is the SDH policy in this thesis.

Previous studies identified treatment and control groups under the slab tax system. In our case, for example, this could have referred to properties around the £500,000 mark, where a higher tax rate would be applied to the total purchase price. This slab structured system led to a bunching effect, where similar properties valued around £500,000 were all transacted below this threshold. As a result, it created a discontinuity in the price distribution and in the amount of tax savings. However, the current stamp duty operates as a progressive tiered system, where tax savings are a continuous function of the TP (see Figure 4.1). If, for instance, one divides the data into multiple groups according to tax savings to perform the DiD, then they will face endogeneity concerns, as the tax saving is simultaneously affected by the price, which in turn is the dependent variable.[24] The worst-case scenario would involve attempting a continuous DiD setting directly with the tax cut rate.

Although it may seem that properties below the SDH threshold are different from those well above the threshold, this does not violate the fundamental principle of identification. The parallel trends assumption can hold conditional on covariates,

[24] The interaction term of treatment groups and the before-after dummy on the right-hand side of a DiD equation is affected by the price-related dependent variables on the left-hand side of the DiD equation

Table 4.6: Control and Treatment Groups

Price	Identification
0 - £125,000	Control group
More than £125,000	Treatment group

provided that a host of covariates, such as property characteristics and location, are controlled for.

It is essential to note that the parallel trends assumption also considers the measurement and transformation of the dependent variable. The assumption of parallel trends requires that the difference in the dependent variables between the treated and control groups remains constant, which may be violated depending on how the difference is measured. Although a logarithmic transformation is often applied to dependent variables in empirical research, it is typically done for the purpose of model interpretation. Nonetheless, correct identification should always take priority when regression modelling is involved.

In the case of DiD, if the parallel trends assumption holds for the dependent variable Y, it may not hold for its logarithmic transformation $log(Y)$ and vice versa. Consider an example where in the pre-treatment period, the outcome Y is 5 for the control group and 10 for the treated group. If, in the counterfactual scenario where treatment never occurred, Y would be 10 for the control group and 15 for the treated group in the post-treatment period, the gap would be $10 - 5 = 5$ before and $15 - 10 = 5$ after, satisfying the parallel trends assumption. However, when considering the logarithmic transformation of Y, $log(Y)$, the gap before treatment would be $log(10) - log(5) = 0.301$, while the gap after treatment would be $log(15) - log(10) = 0.176$. This violates the parallel trends assumption.

Therefore, as the counterfactual is unobserved, the form of the dependent variable should be determined based on its pre-treatment parallel trends test. The forms of the dependent variables used in the subsequent equations are solely for illustrative purposes. If there is any inconsistency between the equations in this section and the results presented in the Results section, the forms of the dependent variables used in the Results section should be considered as the preferred ones.

As mentioned in the Data section, this thesis report models for two categories of dependent variables. Firstly, I aggregate data to compute transaction and listing volumes. Secondly, I utilise transaction-level data, which includes listing prices, transaction prices, price spread, and TOM.

The baseline DiD model for volumes is given as

$$ln(Y_{gt}) = \delta_g + \gamma_t + \beta_0 Treated_g \times After_t + \epsilon_{gt}. \tag{4.10}$$

The dependent variable $ln(Y_{gt})$ is the logarithmic transformed number of monthly transactions of each group – treated and control. The primary parameter of interest is the DiD coefficient β_0, which is interpreted as the effect of the SDH policy on the treated group. This equation is essentially a two-way fixed effect model, where γ_t is the time fixed effect and δ_g is the group fixed effect. $Treated_g$ is an indicator if the observation is in treated group. $After_t$ is an indicator if the transaction(or listing when modelling listing price) happen during SDH. ϵ_{gt} is the corresponding error term.

In addition, I modify above DiD model to allow for a dynamic treatment effect by month. A dynamic DiD shows if the treatment becomes more or less effective over time, or if the effect takes a while to appear. The dynamic DiD model is given as

$$\begin{split} ln(Y_{gt}) = \delta_g + \gamma_t + \phi_{-t_1} Treated + \phi_{-t_1+1} Treated + ...+ \\ \phi_{-1} Treated + \phi_1 Treated + ... + \phi_{t_2} Treated + \epsilon_{gt}. \end{split} \tag{4.11}$$

The model takes June 2020 as the reference time and sets $t = 0$. $t = 1$ would then represent the first month the SDH was implemented, July 2020; $t = 2$ would be the second month after the SDH implementation, and so on. In turn, $t = -1$ implies the second to the last month before the SDH implementation, which is May 2020. The model includes up to t_1 months before the SDH and up to $t = t_2$ months after the SDH implementation. Therefore, $t = -t_1$ indicates the earliest month in the data, which is April 2018.

To avoid perfect multicollinearity during the estimation process, the coefficient for the last period prior to the SDH, ϕ_0, is removed and its effect is incorporated into the constant as the reference point. This is typically accomplished automatically by the statistical modelling software. $\phi_{-t_1}, \phi_{-t_1+1}, \ldots, \phi_{-1}$ are the pre-treatment coefficients and serve as a placebo test to search for an effect before it should exist. These coefficients are expected to be close to zero and not significant; otherwise, it would indicate a violation of the parallel trends assumption. $\phi_1, \phi_2, \ldots, \phi_{t_2}$ are the treatment coefficients, showing the treatment effect month by month; ϕ_1 is the effect one period after treatment, ϕ_2 is the effect two periods after treatment, and so on.

In addition to equation 4.10, a triple DiD (DDD) approach is also used to investigate the heterogeneous effects of the SDH across various property splits. It still follows the same identification strategy with an additional co-variate in the interaction term. The DDD model is given as:

$$\begin{split} ln(Y_{gth}) = \delta_g + \gamma_t + \eta_h + \beta_0 Treated_g \times After_t \times H_h + \beta_1 Treated_g \times H_h + \\ \beta_2 Treated_g \times After_t + \beta_3 After_t \times H_h + \epsilon_{gth}, \end{split} \tag{4.12}$$

where H_h is the heterogeneous factor as an additional co-variate. The factor is a dummy or a categorical variable for two splits of the data (i) by property type[25], (ii) by RUC. Compared with the DiD model in equation 4.10, the DDD model in equation 4.12 simply adds interaction terms related to the heterogeneous factor, $Treated_g \times After_t \times H_h$, $Treated_g \times H_h$, and $After_t \times H_h$. The coefficient of the first term, β_0, is the main interest in this study, and the other two are added for the purpose of model identification. β_0 looks at the effect of SDH on the outcome variable ($Treated_g \times After_t$) and then examines how that effect is different between groups ($Treated_g \times After_t \times H_h$) represented by the heterogeneous factor.

In the previously mentioned model specifications, emphasis was placed on aggregate dependent variables, such as transaction volumes and listing volumes. I will

[25] Previous studies find that the transaction tax has different effects on houses and flats in Germany (Fritzsche and Vandrei 2019; Petkova and Weichenrieder 2017).

now introduce the equivalent model specifications for transaction level dependent variables. The baseline DiD model for transaction data is outlined below:

$$ln(P_{igtcl}) = \beta_0 Treated_g \times After_t + \theta_c + \tau_l + \delta_g + \gamma_t + \boldsymbol{X'}_{igtcl}\boldsymbol{\beta} + \epsilon_{igtcl}. \tag{4.13}$$

The respective dynamic DiD model for the transaction data is given as

$$\begin{aligned} ln(P_{igtcl}) = \phi_{-t_1} Treated_g + \phi_{-t_1+1)} Treated_g + ... + \phi_{-1} Treated_g + \\ \phi_1 Treated_g + ... + \phi_{t_2} Treated_g + \theta_c + \tau_l + \delta_g + \gamma_t + \boldsymbol{X'}_{igtcl}\boldsymbol{\beta} + \epsilon_{igtcl}. \end{aligned} \tag{4.14}$$

The respective DDD model for the transaction data is given as

$$\begin{aligned} ln(P_{igthcl}) = \beta_0 Treated_g \times After_t \times H_h + \beta_1 Treated_g \times H_h + \\ \beta_2 Treated_g \times After_t + \beta_3 After_t \times H_h + \\ \theta_c + \eta_h + \delta_g + \gamma_t + \tau_l + \boldsymbol{X'}_{igthcl}\boldsymbol{\beta} + \epsilon_{igthcl}. \end{aligned} \tag{4.15}$$

P_{igt} represents either (i) the TP, (ii) the initial listing price, (iii) the price spread, or (iv) TOM of property i in group g (g is either control or treatment group) at time t (before or during the SDH). $Treated_g$ is a dummy variable indicating whether the property price is above £125,000, i.e., identifying the treatment group; $After_t$ is a dummy variable denoting a transaction completed during the SDH. θ_c is the CTB for each property. The CTB is exogenous to all market participants, which assists in controlling for unobservable characteristics affecting the dependent variable. All standard errors in estimations are clustered by CTB. τ_l is the location fixed effect at the outcode[26] level; δ_g is the treated or control group fixed-effect; $gamma_t$ represents time fixed effects. $\boldsymbol{X}_{igt}$ is a vector containing other covariates of property

[26] Outcode is short for outward code. It is a smaller area than a local authority, as it comprises the first three digits of the postcode. It includes the postcode area and the postcode district. In this chapter's modelling, there are 1927 unique outcodes in the sample.

characteristics, including total floor area, property built form, property type, number of habitable rooms, number of open fireplaces, current and potential environmental impact, and current and potential energy consumption of the property. β_0 is the coefficient of interest associated with the average price increase in the treated group due to the effect of SDH. Equation 4.15 has an additional variable η_h indicating the heterogeneous factors. I utilise this equation to investigate the heterogeneous effects of SDH with respect to the factors in equation 4.12.

4.6 Results

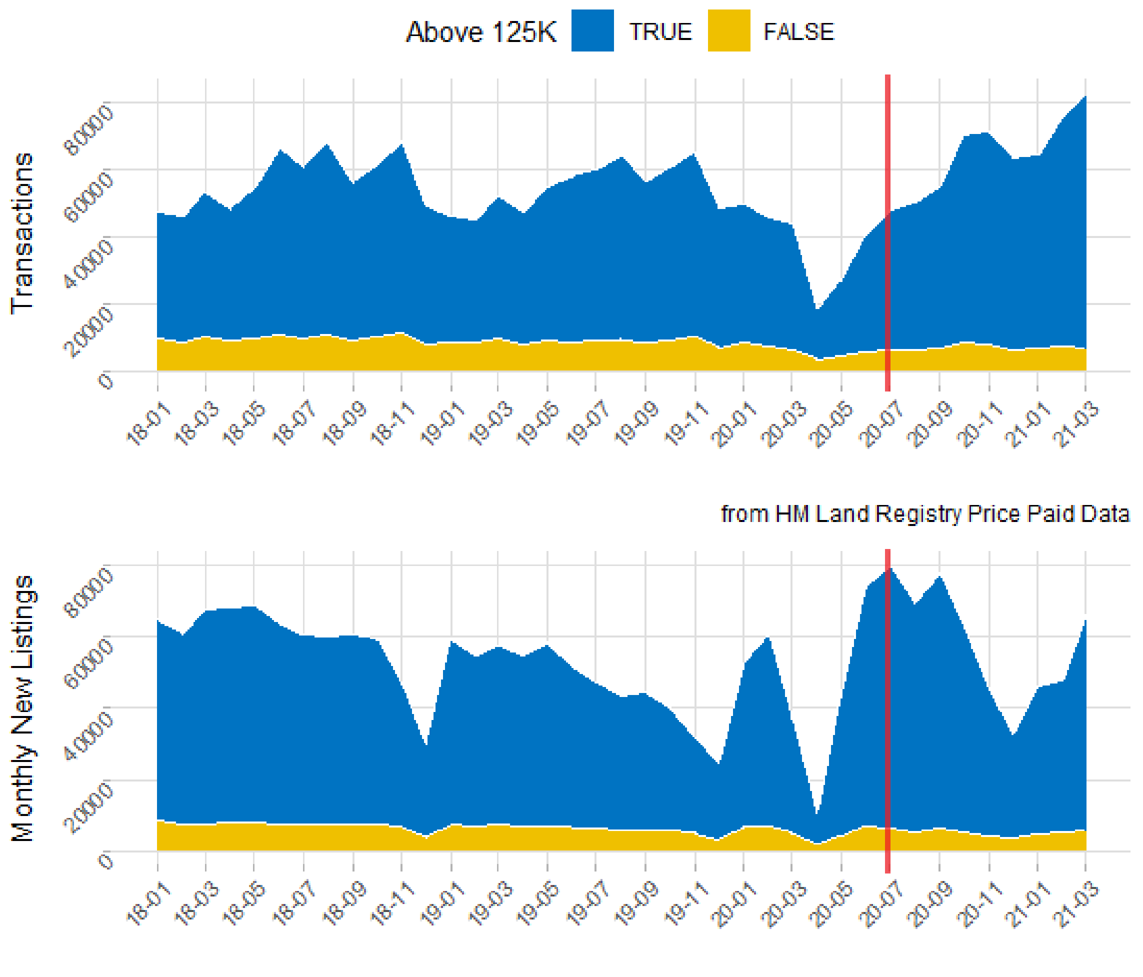
4.6.1 Transactions and Listings

This subsection presents an analysis of both transaction and listing volumes to provide insights into the effects of SDH on supply and demand. The data is acquired by aggregating the monthly number of transactions or listings for both the control group (prices below £125,001) and the treatment group (prices above £125,000). To investigate the dynamic impact of the SDH on various property types, the following steps are undertaken: calculating the monthly number of flat transactions in the control group, determining the monthly number of flat sales in the treatment group, and repeating these two steps for houses. Moreover, the same methodology is employed to examine the dynamic influence of the SDH on properties across different regions, including rural areas, urban cities and towns, and urban conurbations.

The top plot in Figure 4.8 illustrates the full market price property transactions in England from January 2018 to March 2021. Prior to the implementation of the SDH, both the treated and control groups exhibited a similar trend. Nonetheless, following the introduction of the SDH, the trends diverged, with the transaction volume of the treated group continuing to rise until March 2021, the initial expiry date of the SDH, whilst the control group displayed a peak in transactions in October 2020 before declining. In March 2021, there were 82,390 transactions in the treated group, representing an increase of more than 50% compared to the previous years' figures of 44,242, 52,274, and 53,198 in March 2020, March 2019, and March 2018, respectively.

Transactions

Table 4.7 presents the estimations of the DiD model represented in equation 4.10, which examines the effect of SDH on transaction volume. The column 1 demonstrates that, on average, the SDH caused a 53% (calculated as $e^{0.4263} - 1$) increase in transactions for the treated group during the first nine months of the holiday, which was higher than the estimates obtained in studies that analysed the 2008 SDH. Ac-



from HM Land Registry Price Paid Data

collected from Zoopla

Figure 4.8: Monthly Transaction and New Listings from Jan 2018 to March 2021
Note: The red vertical lines mark the onset of the SDH. The sharp drops between March 2020 and June 2020 in both plots were due to the national lockdown during the Covid-19 pandemic. The top plot illustrates the monthly transactions from PPD between January 2018 and March 2021. Prior to the implementation of the SDH, both the treated and control groups exhibited similar trends. Nevertheless, following the introduction of the SDH, the trends diverged. The bottom plot presents the monthly new listings from listing data over the same period. Both groups demonstrated parallel trends with seasonality before the SDH, but the trend of the treated group deviated from the control group during the SDH.

cording to Besley et al. (2014) and Best and Kleven (2017) a 1% reduction in the tax rate in the relevant price window led to a boost effect of 8% to 20% on market activities. With the same calculation, column 2 reveals that the rise in house transactions is 53% and that of flats is 22%. Previous studies on Germany's housing market report that the transaction tax only impacts the sales of single-family houses (Petkova and Weichenrieder 2017).

Table 4.7: SDH Effects on Transaction Volume

	(1) All	(2) Type	(3) RUC
Treated×After	0.4263*** (0.0686)	0.4257*** (0.0730)	0.4059*** (0.0737)
Treated×After×Type:Flat		-0.2215** (0.1032)	
Treated×After×RUC:UrbanCity&Town			-0.0732 (0.1043)
Treated×After×RUC:UrbanConurbation			0.0274 (0.1043)
Num.Obs.	78	156	234
R^2	0.982	0.987	0.987
FE: Treated	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Type		×	
FE: RUC			×

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Only variables of interest are displayed. ”Type” refers to the property type, which is either an apartment/flat or a house; ”RUC” denotes the location type, categorised as ”rural area”, ”urban city and town”, or ”urban conurbation”. The data are generated by counting the number of transactions in the treated and control groups for each month from January 2018 to March 2021 in PPD. For example, in the model represented by column 1, we count the number of transactions in the control group and the treated group for each month, resulting in 78 data points (Number of Observations). Similarly, for the models in columns 2 and 3, we further count the transactions of flats and houses separately, in rural, urban city and town, and urban conurbation areas. Therefore, the Number of Observations is 156 (234). Columns 2 and 3 demonstrate that the SDH has varying impacts with respect to different property types and location types, respectively. The baseline in column 2 is type ”house”, and the baseline in column 3 is ”rural area”. The controlled fixed effects for each model are specified at the bottom of the table.

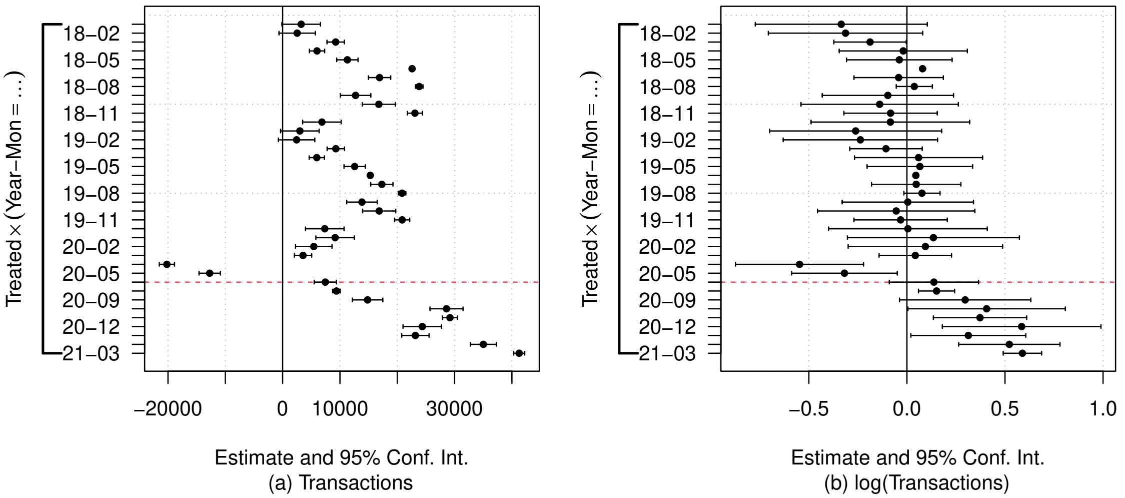
Column 3 demonstrates that there are different effects of SDH on properties located in rural versus urban areas, which could be attributed to market participants taking advantage of SDH to relocate their homes. The average monthly increase caused by SDH in rural areas is 50% (calculated as $e^{0.4059} - 1$). The coefficients of the other two areas are not significantly different from the coefficient of rural area, indicating they have the same level of increases in transactions due to SDH.

According to the UK government data[27], 17.1% of England's population resided in rural areas, 39.5% in urban conurbation areas, and 43.4% in urban city and town areas. Consequently, rural areas experienced the largest relative increase in transactions per capita, while urban conurbations saw the smallest increase per head. Without SDH, one would expect the changes in transaction volumes per head across the three areas to be the same if there were no large household migrations among the areas. The fact that they are not the same suggests that the SDH led to relocation across the three areas, with people utilising the SDH to move from urban conurbation areas to rural areas and urban city and town areas. An urban conurbation area (also referred to as a built-up area) is a region comprising several metropolises, cities, large towns, and other urban areas that, through population growth and physical expansion, have merged to form one continuous urban or industrially developed area. Such areas include Greater London, Greater Manchester, and the West Midlands. They are the polycentric urbanised areas in which transportation has developed to link areas, creating a single urban labour market or travel-to-work areas.

The relocation effect of SDH may be a consequence of the shift to remote work during the pandemic. Jobs remain concentrated in city centres, while workers relocate to the outskirts. Remote work enables workers to transition to telecommuting and enjoy significant welfare benefits, such as reduced commute time and relocation to more affordable neighbourhoods, without sacrificing their desirable jobs (Brueckner et al. 2021; Delventhal et al. 2022).

Although the counterfactual is undoubtedly unobservable, a visual inspection of Figure 4.8 indicates that the parallel trends assumption is not violated in the pre-treatment period. Moreover, a modelling test of the assumption is demonstrated in Figure 4.9. It plots the coefficients $\phi_{-t_1}, \phi_{-(t_1-1)}, ..., \phi_1, \phi_2, ..., \phi_{t_2}$ from equation

[27] 2020 Mid-year population estimates: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1028819/Rural_population__Oct_2021.pdf`



(a) Transactions

(b) log(Transactions)

Figure 4.9: Dynamic Effects and Parallel Trends Test for Transaction Volume
Note: These two plots show the estimated coefficients $(\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2})$ of Equation 4.11 with different forms of the dependent variable. (a) The placebo test of the DiD model with the dependent variable in its original form; (b) The placebo test of the DiD model with the dependent variable in a logarithmic transformation, which is the appropriate choice for DiD in this context.

4.11 associated with the dynamic treatment effects, together with the 95% confidence interval. When the dependent variable is in logarithmic form (Figure 4.9 b), most of the coefficients prior to the SDH implementation are close to zero and insignificant, implying the parallel trends pattern holds before the SDH implementation. The few significant negative coefficients are due to the national lockdown in March 2020, which caused a sharp decline in new search and matching activities. The coefficients since July 2020 show the dynamics of the effect of SDH on transactions. In the UK housing market, most property sales take longer than four weeks. The new search and matching activities caused by the SDH eventually led to a transaction a few months after the listing date. The upward trend indicates stronger effects of the SDH as the time approaches the initially scheduled deadline (31st March 2021). This could be because sellers and buyers are attempting to transact before the deadline to take advantage of the tax reduction. It means that market participants behaved as the policy envisaged, thereby boosting housing activity.

Listings

Regarding the volume of new listings, as illustrated in Figure 4.8 (bottom plot), clear parallel trends are observed in the treated and control groups, with seasonal patterns. The three abrupt drops in market activity prior to the SDH can be attributed to the Christmas holidays and the first national lockdown, which began in late March 2020.

Since the introduction of the SDH, the number of monthly new listings in the treated group has reached a record high level not seen since 2018. Unlike the SDH effect on transactions, there is no lag in time in the volume of newly listed properties. During the SDH, the number of newly listed properties in the treated group (as listed on Zoopla) reached an all-time high in the month the SDH was announced and remained stable for four months. Subsequently, the activity of sellers slowed as the market entered the 2020-21 Christmas holidays, resulting in a slightly lower supply level in January and February 2021, as the market anticipated the end of SDH. However, the trend rebounded immediately in March 2021 upon the government's announcement of the extension of the SDH.

Table 4.8 illustrates a positive and significant impact of SDH on the monthly number of new listings (as shown in column 1). The coefficient of 0.4739 suggests, on average, that the SDH increased the number of monthly new listings by 60%. This finding is novel and has not been documented in previous studies. The boosting effect is evident for both houses and flats, with the increase for flats being slightly higher, but not significantly different from that of houses.

Column 3 demonstrates the varied impact of the SDH across urban and rural areas. It led to a 38% increase in monthly new listings in rural areas. Although the coefficient of urban conurbation area is slightly greater, and the coefficient of urban city and town areas is slightly lower, these differences are not statistically significant when compared to the coefficient of rural areas.

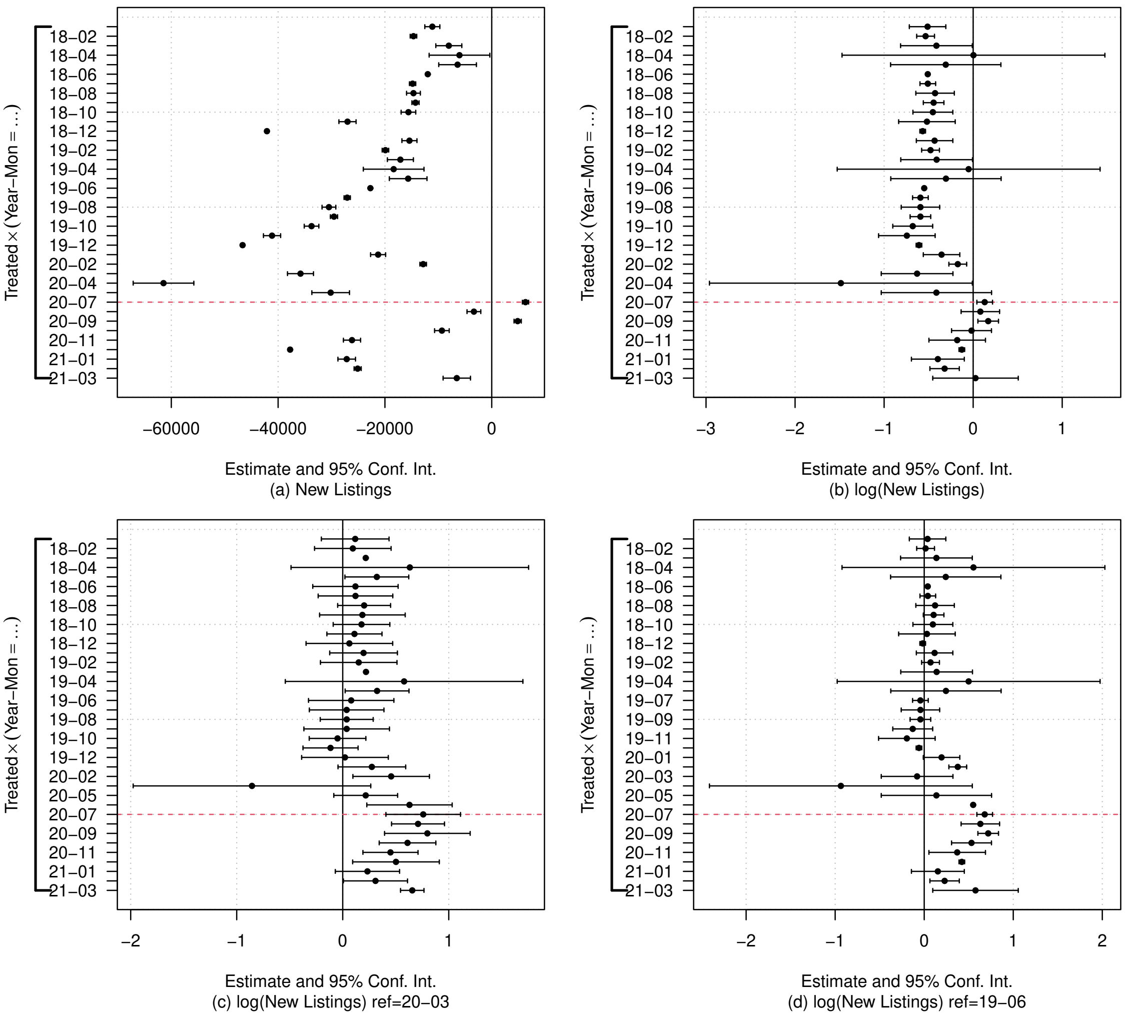
While the parallel trends assumption is upheld in the pre-SDH period through visual inspection of the lower plot in Figure 4.8, the placebo test in Panel (b) of Figure 4.10 using June 2020 as a reference does not provide conclusive evidence. The

Table 4.8: SDH Effect on Monthly New Listings

	(1) All	(2) Type	(3) RUC
Treated×After	0.4739***	0.3159***	0.3539***
	(0.1140)	(0.1172)	(0.1159)
Treated×After×Type:Flat		0.0552	
		(0.1657)	
Treated×After×RUC:UrbanCity&Town			-0.0805
			(0.1639)
Treated×After×RUC:UrbanConurbation			0.0234
			(0.1639)
Num.Obs.	78	156	234
R^2	0.966	0.971	0.975
FE: Treated	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Type		×	
FE: RUC			×

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Only variables of interest are displayed. The data is generated by counting the new listings in treated and control groups for each month from listing data. ”Type” refers to the property type, which is either an apartment/flat or a house; ”RUC” refers to the location type, being either a ”rural area”, ”urban city and town”, or ”urban conurbation”. Columns 2 and 3 demonstrate that the SDH has varying impacts with respect to different property types and location types, respectively. The baseline in column 2 is the type ”house”, and the baseline in column 3 is ”rural area”. The controlled fixed effects for each model are specified at the bottom of the table.



(a) New Listings

(b) log(New Listings)

(c) log(New Listings) ref=20−03

(d) log(New Listings) ref=19−06

Figure 4.10: Dynamic Effects and Parallel Trends Test for Monthly New Listings *Note:* These plots show the estimated coefficients $(\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2})$ of Equation 4.14 with different forms of the dependent variable or with different reference time points. (a) the placebo test of the DiD model with the dependent variable in its original form; (b) the placebo test of the DiD model with the dependent variable in a logarithmic transformation, which is the appropriate choice for DiD in this context, the reference time being June 2020; Panels (c) and (d) present the placebo tests with reference times being March 2020 and June 2019, respectively.

pre-SDH coefficients are largely negative and significant, around -0.5. This could potentially be due to an overcompensation in new listings in June, as it was the first month when most restrictions were lifted since the national lockdown. This resulted in June 2020 new listings being significantly higher than in the same month in 2018 and 2019. The test results provide unambiguous evidence supporting the parallel trends assumption in the pre-treatment period when considering March 2020 and June 2019 as reference times in Panels (c) and (d) of Figure 4.10, respectively. Additionally, a timing effect in market supply is observed, with fewer potential sellers entering the market due to the odds of finding a buyer and completing before the tax holiday's expiration. The market did not anticipate the SDH extension announced in March 2021, leading to a slowdown in new listings four months before the expected deadline.

4.6.2 Prices

Transaction Prices

Table 4.9 presents the modelling outcomes for transaction prices based on two distinct samples. The "Baseline Sample" section illustrates estimations derived from a multi-source integrated big dataset, which provides a more comprehensive range of control variables, albeit with a smaller sample size. Conversely, the "Large Sample" section encompasses 92% of the population's transactions from the PPD, incorporating variables from the EPC and CTB datasets, but excludes several control variables that are solely available in listing data. The results from both the "Baseline Sample" and "Large Sample" sections exhibit consistency, indicating that the "Baseline Sample" delivers robust population estimations despite its reduced sample size.

On average, the SDH led to an approximately 2% (1.8% from baseline sample and 2.5% from large sample) increase in the TP. The average TP for the treated group in the population (based on the PPD) was £384,914. In a counterfactual scenario without the SDH, the average TP would have been £377,367.[28] The stamp duty for the average property in this counterfactual scenario would have amounted to £3,868

[28] Calculated as $384,914/(1+0.02)$.

Table 4.9: SDH Effects on Transaction Price

Depend.	(1) TP	(2) TP(Type)	(3) TP(RUC)
Baseline Sample (multi-source): N=814,937			
Treated×After	0.0184***	0.0196**	0.0254***
	(0.0015)	(0.0060)	(0.0071)
Treated×After×Type:Flat		-0.0052	
		(0.0075)	
Treated×After×RUC:UrbanCity&Town			-0.0084***
			(0.0020)
Treated×After×RUC:UrbanConurbation			0.0039
			(0.0031)
R^2	0.930	0.931	0.931
Large Sample: N=2,142,189			
Treated×After	0.0247***	0.0257**	0.0368**
	(0.0026)	(0.0105)	(0.0108)
Treated×After×Type:Flat		-0.0111	
		(0.0068)	
Treated×After×RUC:UrbanCity&Town			-0.0056**
			(0.0021)
Treated×After×RUC:UrbanConurbation			-0.0167***
			(0.0023)
R^2	0.912	0.912	0.912
FE: CTB	×	×	×
FE: Outcode	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Treated	×	×	×
FE: Property Type		×	
FE: RUC			×

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: The table only displays variables of interest. The ”Baseline Sample” section presents the regression outcomes using the baseline sample (Listings-PPD-EPC-CTB), wherein the sample size is reduced to accommodate a more extensive range of control variables during the data matching process. The ”Large Sample” section exhibits modelling results from a larger sample (PPD-EPC-CTB), comprising 92% of the population transactions recorded in PPD. This table demonstrates that the ’Baseline Sample’ yields robust results representative of the population. ”Type” refers to the property type, which can be either an apartment/flat or a house; ”RUC” denotes the location type, categorised as rural, urban city and town, or urban conurbation. Columns 2 and 3 display regression results including controls for property type and RUC, respectively. The baseline in column 2 is a house, and the baseline in column 3 is a rural area. The controlled fixed effects for each model are listed at the bottom. All standard errors are clustered by CTB.

for first-time buyers and £8,868 for home movers replacing their main residence.[29] Consequently, the average TP increase due to SDH is approximately twice the amount of SDLT that first-time buyers would have paid if the SDH had not been in place. For buyers replacing their main residence, the increase in TP accounts for roughly 85% of the tax-saving. These findings significantly exceed those reported in previous studies. Besley et al. (2014) demonstrate that sellers only received 40% of the tax saving from the 2008 SDH. If the impact of taxation and tax deductions on the property market were symmetrical, the economic burden of the new slice stamp duty tax would fall almost entirely on the sellers, in line with the conclusions of Dachis et al. (2012) and Davidoff and Leigh (2013).

In column 2 of the baseline, the coefficient of "Treated×After" represents the effect of SDH on house transactions. It shows similar estimates to those obtained from the large sample. It implies that on average, the SDH increases house prices by 2%, whereas prices increase for apartments is smaller at 1.5% but not significantly different from that of houses.

The average transaction prices of houses and apartments in the SDH-affected group are £384,261 and £388,905, respectively. Without the SDH, the prices for houses and apartments would have been £376,727 and £383,158, respectively.[30] The effective stamp duty rate for home movers would then be 2.35%[31] and 2.39%[32] for houses and apartments, respectively. For first-time buyers, the rate would be 1.02%[33] and 1.09%[34] for houses and apartments, respectively. Consequently, on average, first-time buyers would have faced a considerably lower total cost if the SDH had not been implemented.

Column 3 demonstrates that the SDH resulted in an approximate 2.6% increase in transaction prices in rural areas, 1.7% in urban city and town areas, and 3.0% in urban conurbation areas. The average TP for the treated group in these three areas

[29] Based on the tax rate in table 4.1.

[30] Calculated as $384261/(1 + 0.02)$ and $388905/(1 + 0.015)$, respectively.

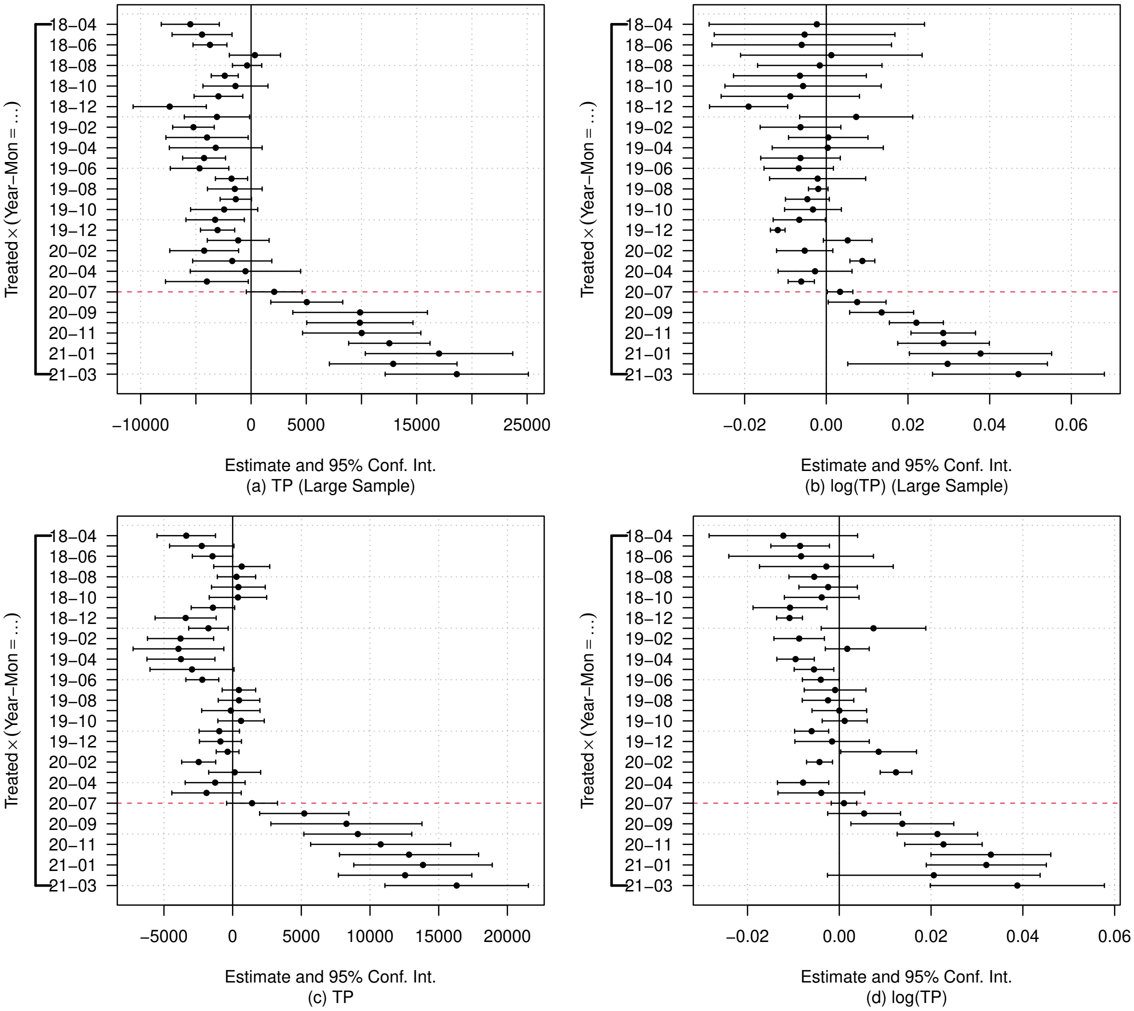
[31] $((376727 - 250000) \times 0.05 + 125000 \times 0.02)/376727$

[32] $((383158 - 250000) \times 0.05 + 125000 \times 0.02)/383158$

[33] $(376727 - 300000) \times 0.05/376727$

[34] $(383158 - 300000) \times 0.05/383158$

amounts to £399,841, £333,057, and £453,494, respectively. Employing the same calculation method, the effective stamp duty rate would be 2.4%, 1.9%, and 2.7% for home movers in the respective areas. For first-time buyers, the rate would be 1.2%, 0.4%, and 1.6%, respectively. On average, only home movers relocating to urban city and town areas experience a reduction in expenditure because of SDH.



(a) TP (Large Sample)

(b) log(TP) (Large Sample)

(c) TP

(d) log(TP)

Figure 4.11: Dynamic Effects and Parallel Trends Test for Transaction Price
Note: These plots show the estimated coefficients $(\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2})$ of Equation 4.14 with different forms of the dependent variable or with different samples. (a) and (b) represent the placebo tests of the DiD model, with the dependent variable in its original form or logarithmic transformation, using the large sample. (c) and (d) display the same tests with the baseline sample. Both sets of tests suggest that the logarithmic transformation of the outcome variable is the appropriate choice for the DiD analysis in this context. The reference time for all tests is June 2020.

A proper visual inspection method to verify the parallel trends assumption for

DiD with property-level micro data is not available. Instead, a placebo test with equation 4.14 is presented. This test follows the same intuition as the examination for transaction and listing volumes in the previous subsection and maintains a similar model specification. Figure 4.11 displays the parallel trends test results. Panels (a) and (b) present results using the large sample, while (c) and (d) show test results based on the baseline sample. Compared to panels (a) and (c), panels (b) and (d) provide stronger evidence supporting parallel trends in the pre-SDH period, indicating that the logarithmic transformed dependent variable is an appropriate choice for DiD in this context. Additionally, panels (b) and (d) reveal that the price-raising effect of SDH intensified as time approached the initial deadline.

Initial Listing Prices

I employ the same models 4.13 and 4.15 to estimate the SDH's effects on the initial asking price and test the parallel trends assumption with equation 4.14. Similar to the previous analysis, Table 4.10 demonstrates that the results from the large sample and our baseline sample estimation are consistent. On average, the listing price increased by 2.1-3%, which is, on average, 0.5% larger than the increase in transaction prices. The impact of SDH on the initial listing price is novel and has not been previously documented in the literature. This finding aligns with the data presented in the Rightmove Asking Price Index depicted in Figure 4.4. The subsequent interpretation of the influence of SDH on various property types and RUC areas is based on the estimates provided in the baseline sample section.

Column 3 illustrates that the SDH led to a 3.5% to 4.6% increase in the initial listing price for houses, while there was minimal or no change in the initial listing price for apartments. The rise in asking price for houses is significantly larger than the increase in the TP, suggesting that sellers were aware of the heightened demand for houses and adjusted their expectations accordingly.

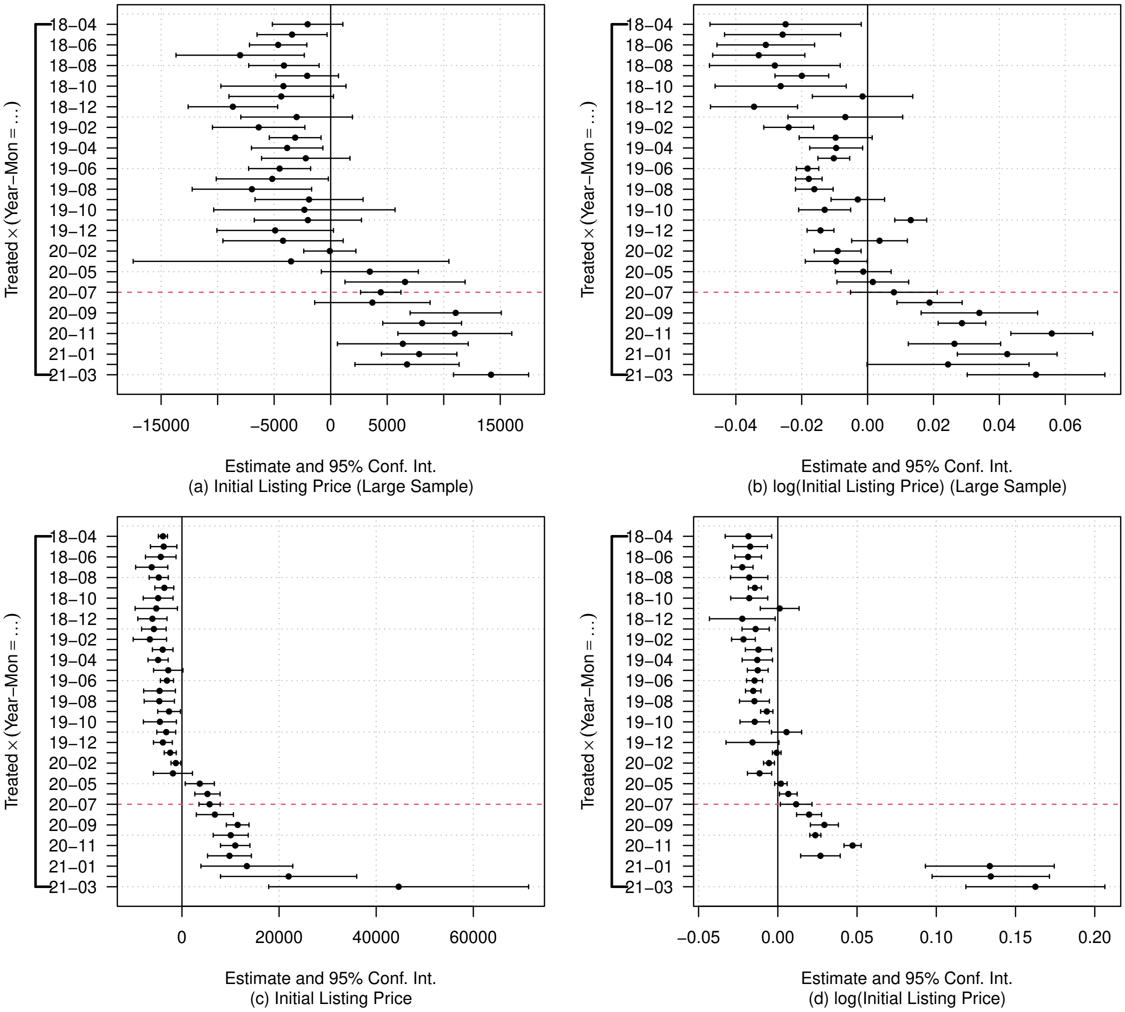
Column 4 reveals that, due to SDH, the initial listing price increased considerably more in rural areas and urban conurbation areas by over 4.4% and 4.8% respectively, with a relatively modest increase of 2.5% observed in urban city and town areas.

Table 4.10: SDH Effects on Initial Listing Price

Depend.	(1) LP	(2) LP(Type)	(3) LP(RUC)
Baseline Sample (multi-source): N=675,701			
Treated×After	0.0219***	0.0347***	0.0428***
	(0.0022)	(0.0075)	(0.0046)
Treated×After×Type:Flat		-0.0256	
		(0.0153)	
Treated×After×RUC:UrbanCity&Town			-0.0185***
			(0.0024)
Treated×After×RUC:UrbanConurbation			0.0044
			(0.0071)
R^2	0.929	0.929	0.929
Large Sample: N=1,349,352			
Treated×After	0.0292***	0.0449**	0.0633**
	(0.0028)	(0.0131)	(0.0192)
Treated×After×Type:Flat		-0.0484***	
		(0.0045)	
Treated×After×RUC:UrbanCity&Town			-0.0296**
			(0.0119)
Treated×After×RUC:UrbanConurbation			-0.0228*
			(0.0100)
R^2	0.909	0.909	0.909
FE: CTB	×	×	×
FE: Outcode	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Treated	×	×	×
FE: Property Type		×	
FE: RUC			×

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: The table only displays variables of interest. The ”Baseline Sample” section presents the regression outcomes using the baseline sample (Listings-PPD-EPC-CTB), wherein the sample size is reduced to accommodate a more extensive range of control variables during the data matching process. The ”Large Sample” section exhibits modelling results from a larger sample (PPD-EPC-CTB), comprising 92% of the population transactions recorded in PPD. This table demonstrates that the ’Baseline Sample’ yields robust results representative of the population. ”Type” refers to the property type, which can be either an apartment/flat or a house; ”RUC” denotes the location type, categorised as rural, urban city and town, or urban conurbation. Columns 2 and 3 display regression results including controls for property type and RUC, respectively. The baseline in column 2 is a house, and the baseline in column 3 is a rural area. The controlled fixed effects for each model are listed at the bottom. All standard errors are clustered by CTB.



(a) Initial Listing Price (Large Sample)

(b) log(Initial Listing Price) (Large Sample)

(c) Initial Listing Price

(d) log(Initial Listing Price)

Figure 4.12: Dynamic Effects and Parallel Trends Test for Initial Listing Price
Note: These plots show the estimated coefficients $(\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2})$ of Equation 4.14 with different forms of the dependent variable or with different samples. Panels (a) and (b) represent the placebo tests of the DiD model with the dependent variable in original form or logarithmic transformation using the large sample. Panels (c) and (d) display the same tests with the baseline sample. The reference time for all tests is June 2020.

The parallel trends tests from the corresponding dynamic DiD model are displayed in Figure 4.12. In comparison to panels (a) and (c), panels (b) and (d) exhibit a better alignment with the pre-treatment parallel trends, particularly in the year preceding the SDH. Consequently, the log-transformed initial listing price is chosen as the dependent variable in the DiD modelling. The tests also uncover a trend in which the increase in the initial listing price due to SDH became more pronounced as the holiday deadline approached.

4.6.3 Liquidity: Price Spread and Time on Market

Past literature has not explicitly examined the impact of property transaction tax on market liquidity. This aspect is analysed using the same identification strategy employed in the previous price analysis. The same equations (4.13, 4.15, and 4.14) are adopted, with the dependent variable being either the price spread or TOM. When modelling for price spreads, I include the final asking price as an additional control variable. This approach helps investigate the impact of SDH on liquidity without confounding the prices. For instance, it eliminates the situation of outlier price spreads where sellers attempt their luck, set a high asking price, and then transact at the market price later, which establishes a correlation suggesting that a high asking price might lead to a high price spread. Similarly, when modelling for TOM, I incorporate the initial listing price as an additional control variable. As the results from both the large sample and the baseline sample are consistent, as demonstrated in previous subsections, we can directly interpret the results in this section in conjunction with previous findings.

The price spread is measured as the difference between the final listing price and the TP. Utilising the final asking price offers a better measure of the spread, as some sellers may initially overprice their properties intentionally to test the market and later revise the price downwards gradually (Huang and Milcheva 2020); this occurred in 35% of the cases in my sample.

The average price spread is positive, indicating that the final listing price is, on average, higher than the TP. Considering that the majority of the SDH savings are passed onto sellers and the TP continues to rise, it is expected that the price spread would decrease during the SDH, as buyers possess less bargaining power. Table 4.11 presents results consistent with the expectation. On average, the price spread narrows by $£2,703$ significantly due to the SDH and continues to become smaller as the SDH nears its conclusion. The effect is more pronounced for houses than for flats, as illustrated in Figure 4.13.

Table 4.11: SDH Effects on Price Spreads

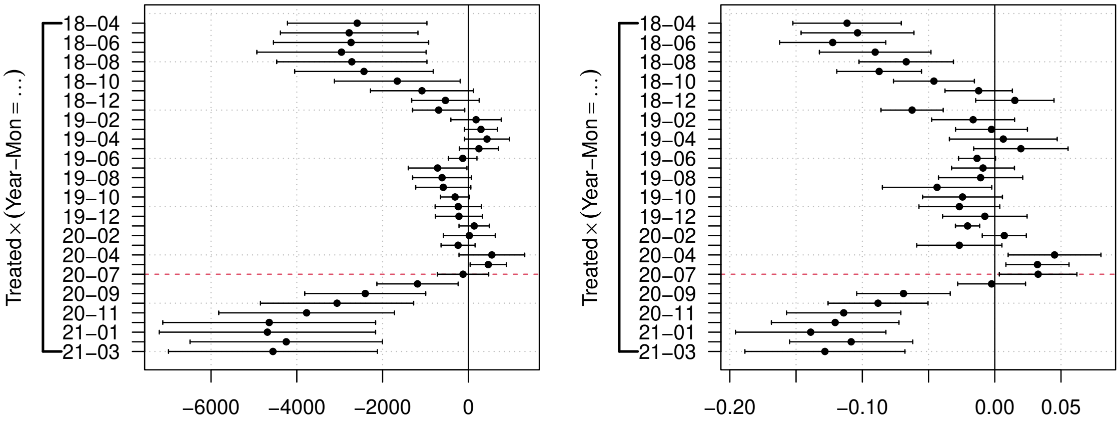
Depend.	(1) PS	(2) PS(Type)	(3) PS(RUC)
Baseline Sample (multi-source): N=814,937			
Treated×After	-2703.29***	-2394.71**	-2408.28**
	(721.91)	(692.43)	(719.35)
Treated×After×Type:Flat		1734.67	
		(1129.25)	
Treated×After×RUC:UrbanCity&Town			629.68
			(449.89)
Treated×After×RUC:UrbanConurbation			-800.61
			(425.03)
R^2	0.277	0.278	0.277
FE: CTB	×	×	×
FE: Outcode	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Treated	×	×	×
FE: Property Type		×	
FE: RUC			×

$* p < 0.1$, $** p < 0.05$, $*** p < 0.01$

Note: The table displays only the variables of interest. The baseline sample is used for modelling, as it requires matching all sources to calculate the price spread. "Type" refers to the property type, which can be either an apartment/flat or a house; "RUC" denotes the location type, categorised as rural, urban city and town, or urban conurbation. Columns 2 and 3 display regression results including controls for property type and RUC, respectively. The baseline in column 2 is a house, and the baseline in column 3 is a rural area. The controlled fixed effects for each model are listed at the bottom. All standard errors are clustered by CTB.

Considering that both sellers and buyers endeavour to complete transactions before the SDH expires, and taking into account the final listing price, the outcome can be construed as buyers being prepared to propose a price that satisfies the sellers' reservation price. In turn, sellers remain unwilling to reduce their asking prices. In effect, buyers acquiesce to paying higher prices than they would in the absence of the SDH in order to capitalise on the "advantage" offered by the SDH. Such behaviour was not observed prior to the implementation of the SDH, as demonstrated in Huang and Milcheva (2020). This indicates that agents modify their pricing and negotiation strategies in response to the policy.

The impacts of the SDH on the price spread can be expected to vary for houses and flats. The price spread is likely to be narrower for houses compared to flats during the SDH period. This can be attributed to the higher demand for houses relative to flats, which results in a more robust bargaining position for house sellers. Evidence from earlier sections reveals that, although the increase in new listings is similar for both houses and flats, the SDH has augmented the transaction volume of houses to a greater extent than that of flats, indicating a stronger demand for houses. Column 2 in Table 4.11 substantiates this hypothesis.



Estimate and 95% Conf. Int.
(a) Price Spread

Estimate and 95% Conf. Int.
(b) log(Price Spread)

Figure 4.13: Dynamic Effects and Parallel Trends Test for Price Spreads
Note: These plots show the estimated coefficients ($\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2}$) of Equation 4.14 with different forms of the dependent variable. (a) and (b) represent the placebo test of the DiD model, with the dependent variable in its original form and logarithmic transformation, respectively. In this context, the original form of the outcome variable is the suitable choice for the DiD approach. The reference time for all tests is June 2020.

As demonstrated in previous findings in this chapter, rural areas, in contrast to urban areas, experienced the most significant increase in listing and transaction prices, as well as transaction volume, owing to the SDH. This highlights a strong demand for properties in rural locations. In light of the discussion on bargaining behaviour, it can be expected that transactions in rural areas would display the narrowest price spread. This assertion is supported by column (3) in Table 4.11.

Table 4.12: SDH Effects on TOM

Depend.	(1) TOM	(2) TOM(Type)	(3) TOM(RUC)
Baseline Sample (multi-source): N=814,937			
Treated×After	-4.63	4.21*	-2.85
	(3.11)	(1.83)	(5.20)
Treated×After×Type:Flat		-16.91**	
		(6.03)	
Treated×After×RUC:UrbanCity&Town			4.36
			(2.85)
Treated×After×RUC:UrbanConurbation			-3.08
			(5.63)
R^2	0.096	0.097	0.098
FE: CTB	×	×	×
FE: Outcode	×	×	×
FE: Year	×	×	×
FE: Month	×	×	×
FE: Treated	×	×	×
FE: Property Type		×	
FE: RUC			×

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays only the variables of interest. The baseline sample is used for modelling, as it requires matching all sources to calculate the TOM. ”Type” refers to the property type, which can be either an apartment/flat or a house; ”RUC” denotes the location type, categorised as rural, urban city and town, or urban conurbation. Columns 2 and 3 display regression results including controls for property type, and RUC, respectively. The baseline in column 2 is a house, and the baseline in column 3 is a rural area. The controlled fixed effects for each model are listed at the bottom. All standard errors are clustered by CTB.

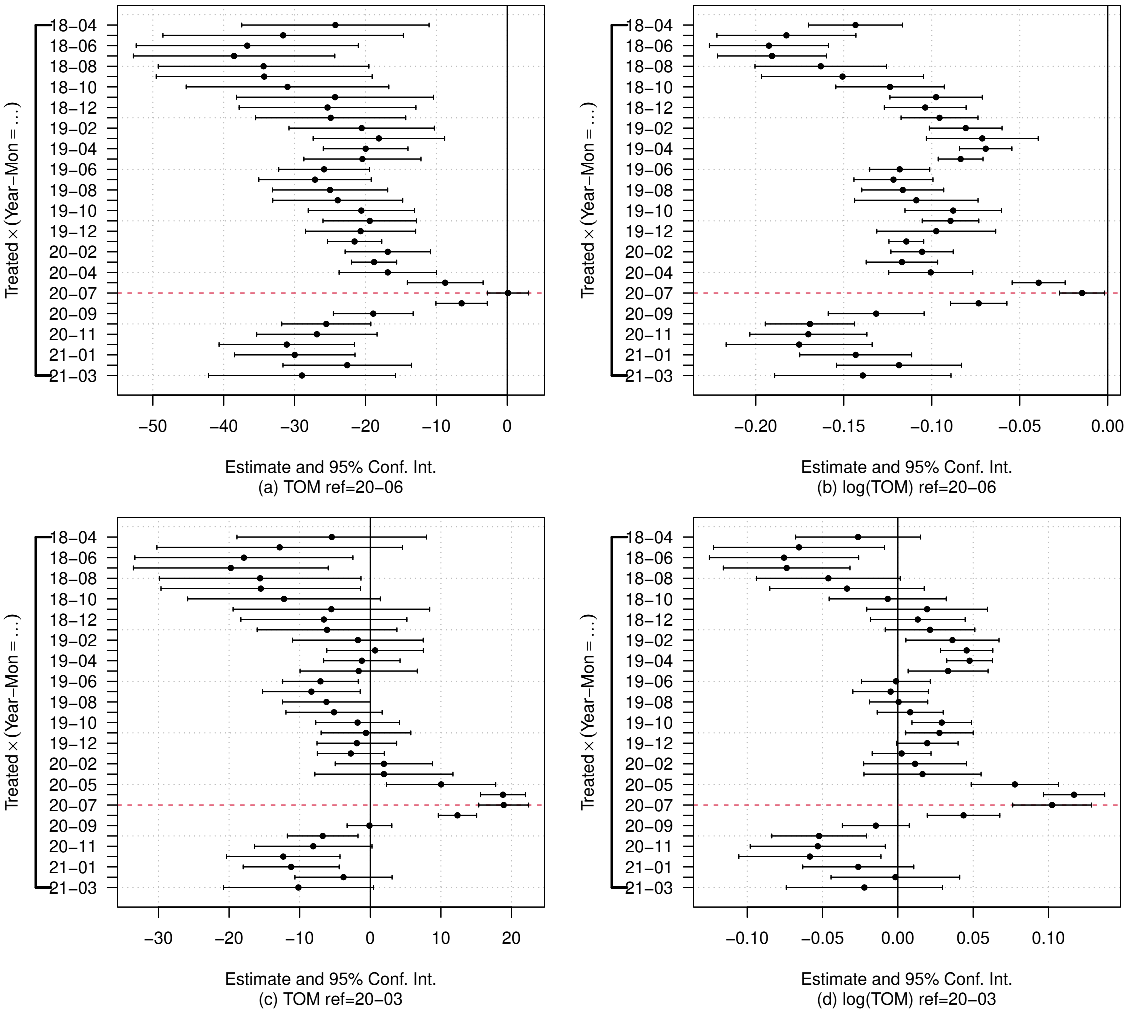
The analytical results pertaining to the impact of the SDH on the TOM revealed only a minor and statistically insignificant negative effect. On average, transactions subject to SDH tax savings spent 4.6 fewer days on the market compared to the duration they would have experienced without the SDH, as indicated in column 1 of Table 4.12. According to a report from LSE London for Family Building Society[35], the SDH exerted considerable pressure on the conveyancing system, leading to an

[35] Lessons from the stamp duty holiday: `https://www.lse.ac.uk/geography-and-environment/research/lse-london/documents/Reports/Lessons-from-stamp-duty-holiday-LSE-London-Report-2021.pdf`

increase in conveyancing time from 12 weeks to nearly 16 weeks, with buyers facing difficulties in finding solicitors. Furthermore, the average processing time for mortgage applications also increased from 2 weeks to 4 weeks. These factors could potentially contribute to an increase in the TOM for transactions. Despite the SDH effectively stimulating the housing market, its impact on TOM was not found to be substantial or significant.

Nonetheless, when examining the heterogeneous effects of the SDH on the TOM for houses and flats, it is anticipated that the SDH would exert a more pronounced negative impact on flats, considering that the demand (transaction volume) for flats is considerably lower than that for houses. This could be attributed to the fact that 92% of the flats in my sample are situated in urban areas, which exhibit high population density, rendering them less appealing during the Covid-19 pandemic. Moreover, 43% of flat sales are chain-free, compared to a mere 31% of house sales. This makes a flat transaction less likely to fall through and is associated with shorter TOM. The chain-free indicator also implies that these units might be buy-to-let properties released for sale by investors seeking to liquidate their assets and capitalise on the SDH. Consequently, the SDH differentiates the demands for houses and flats. Column 2 confirms my assumptions that the SDH has a positive effect on the TOM for houses but a significantly negative effect on flats. On average, the policy led to houses spending 4.2 more days on the market, while flats spent 16.9 fewer days compared to the case of the houses. Furthermore, I find no significant difference in the impact of the SDH on the TOM across RUC areas, as illustrated in column 3.

Figure 4.14 presents the placebo tests. Panels (a) and (b) are tests with the TOM and $log(TOM)$, respectively, using June 2020 as the reference point. As illustrated in Figure 4.10, the nationwide lockdown caused the TOM in June 2020 to be significantly longer than under normal circumstances, leading to numerous negative coefficients in the test. Panels (c) and (d) represent the same tests with March 2020 as the reference point, which is the closest month to July 2020 and remained unaffected by the lockdown in terms of TOM measurement. These panels suggest that the original form of the dependent variable is the most appropriate for the DiD analysis in this



(a) TOM ref=20−06

(b) log(TOM) ref=20−06

(c) TOM ref=20−03

(d) log(TOM) ref=20−03

Figure 4.14: Dynamic Effects on TOM

Note: These plots show the estimated coefficients $(\phi_{-t_1}, \ldots, \phi_{-1}, \phi_1, \ldots, \phi_{t_2})$ of Equation 4.14 with different forms of the dependent variable or with different reference time points. (a) and (b) represent the placebo tests of the DiD model with the dependent variable in its original form and in logarithmic transformation, respectively, using June 2020 as the reference time. Similar to the issue depicted in Figure 4.10, the national lockdown led to an extended TOM in June 2020 compared to typical circumstances, resulting in most coefficients in the test being negative. (c) and (d) comprise the same tests with March 2020 as the reference time, which is the closest month to July 2020 and was not influenced by the lockdown in terms of measuring the TOM. These tests demonstrate that the original form of the outcome variable is the suitable choice for the DiD analysis in this context.

context.

4.7 Conclusion

Prior studies (Besley et al. 2014; Best and Kleven 2017; Hilber and Lyytikäinen 2017) have explored the UK transaction taxation under the former slab system, which was in place until 2014. Nevertheless, there is limited knowledge regarding the impact of the new progressive tiered stamp duty on the market. This chapter seeks to bridge this gap by conducting a comprehensive investigation of the effects of transaction tax on the UK residential market, utilising the 2020 SDH as a quasi-natural experiment. Implemented in the UK in July 2020 amid the initial stages of the Covid-19 pandemic, the 2020 SDH is a property transaction tax reduction policy. Its objectives were to stimulate market activity and the economy by: (1) lowering acquisition costs, (2) encouraging buyers to conduct transactions during an economic downturn, and (3) providing additional funds for home movers to allocate towards moving-related goods and services. My findings corroborate the second objective, but contradict the first and third objectives.

My findings reveal that the SDH significantly influenced housing market activity, resulting in a considerable increase of 60% in supply and 53% in transactions during the most challenging phase of the Covid outbreak. Moreover, the SDH contributed to a rise in both asking and transaction prices. The average transaction price increase ranged from approximately 1.9-2.5%, whilst the average initial listing price increase was between 2.1-3%. This suggests that the entire tax savings arising from the SDH were passed on to sellers through increased prices. First-time buyers, who were expected to benefit from the SDH, ended up paying over twice the amount of the stamp duty when purchasing an average property, had the SDH not been in place, while home movers paid more than 85% of the tax savings. The escalation in both transactions and prices could potentially lead to heightened consumption of housing-related services. As per Best and Kleven (2017), moving homes incurs additional spending by movers, amounting to roughly 5% of the home's value. For homeowners who are not relocating, the price increase might also stimulate an upsurge in regional consumption. Campbell and Cocco (2007) demonstrate that house price increases are most likely to benefit older homeowners and enhance consumption by making households feel wealthier or by relaxing borrowing constraints.

Furthermore, this chapter demonstrates that an unanticipated temporary removal of transaction taxes results in timing behaviour by market agents. The deadline for the SDH was announced simultaneously with the SDH launch, enabling agents to fully anticipate it and adjust their behaviour when engaging in the market. The SDH impact on monthly new listings transitions from a stimulant at the beginning of the tax holiday to a deterrent four months before the deadline, dissuading potential sellers from entering the market due to the reduced likelihood of finding a buyer and finalising deals before the tax holiday expires.

I find that the positive spread between the final asking price and the TP significantly diminished, indicating a narrowing as the deadline approached. This suggests that sellers held a stronger bargaining position, and buyers were willing to increase their offers to meet sellers' demands in order to capitalise on the SDH.

Additionally, I find evidence that market participants utilise the SDH to relocate away from highly urbanised polycentric areas. Residing in these areas provides excellent job accessibility and an associated price premium for households. However, the option to work from home has enabled workers to switch to telecommuting, yielding significant welfare gains for individuals by saving commute time. This could have prompted a shift in demand from apartments to houses during the pandemic, as people sought more affordable neighbourhoods. This trend is reinforced by the observation that the SDH led to a shift in demand from apartments to houses. As most apartments in England are situated in dense urban areas, they become less appealing purchases when working from home is a viable option for a greater number of individuals.

This thesis examines the impact of SDH on all properties eligible for tax reduction, introducing a potential challenge in directly comparing the control and treated groups. Consequently, there exists a limitation wherein the estimated effect may be influenced by unobserved factors. Future research endeavours could focus on exploring the average treatment effect of the policy on specific groups. This can be achieved by defining subsamples to ensure that both control and treatment groups encompass a more homogeneous set of properties. This approach would help mitigate the impact of varying house price dynamics.

Proof of Equation 4.2

The price maximisation equation is given as:

$$\underset{p}{\operatorname{argmax}}(b + K - (1 + \tau)p)^{\alpha}(p - s)^{1-\alpha} \tag{4.16}$$

where τ and K are fixed numbers when the price falls in a certain band. Considering this equation is a function of p, take the logarithm of the function we get:

$$\alpha \ln (b + K - (1 + \tau)p) + (1 - \alpha) \ln (p - s) \tag{4.17}$$

Then by the first order condition (taking the derivative with respect to p and set it equals to zero):

$$-\frac{\alpha(1 + \tau)}{b + K - (1 + \tau)p} + \frac{(1 - \alpha)}{(p - s)} = 0 \tag{4.18}$$

Then solving above equation yields the following formula for the price:

$$p = \frac{1 - \alpha}{1 + \tau}(b + K) + \alpha s \tag{4.19}$$

□

Proof that Seller's Surplus Increases during SDH

Surplus before SDH:

$$S^S{}_0 = (1 - \alpha_0)(\frac{b_0 + K_0}{1 + \tau_0} - s_0) \tag{4.20}$$

Surplus during SDH:

$$S^S{}_1 = (1 - \alpha_1)(\frac{b_1 + K_1}{1 + \tau_1} - s_1) \tag{4.21}$$

Assumption 1: $\alpha_1 < \alpha_0$.

Assumption 2: $s_1 = s_0$.

Assumption 3: $b_1 \geq b_0$.

Then the change in the seller's surplus is: $\Delta S^S = (17) - (16) = s_0(\alpha_0 - \alpha_1) + [(1 - \alpha_1)\frac{b_1 + K_1}{1 + \tau_1} - (1 - \alpha_0)\frac{b_0 + K_0}{1 + \tau_0}]$. The $s_0(\alpha_0 - \alpha_1)$ is always positive due to assumption

1. I only need to check the sign of the second part, $(1-\alpha_1)\frac{b_1+K_1}{1+\tau_1} - (1-\alpha_0)\frac{b_0+K_0}{1+\tau_0}$, denoted as ΔS_{p2}.

According to Table 4.2 and above assumptions:

For $p <= £125,000$: $\Delta S_{p2} = (1-\alpha_1)b_1 - (1-\alpha_0)b_2 > 0$.

For $£125,000 < p <= £250,000$, I have $b_1 >= p > 125,000$, then:

$$\begin{aligned}\Delta S_{p2} &= (1-\alpha_1)b_1 - (1-\alpha_0)\frac{b_0+2500}{1+0.02} \\ &= \frac{[(1-\alpha_1)b_1 - (1-\alpha_0)b_0] + [(1-\alpha_1)0.02b_1 - (1-\alpha_0)2500]}{1.02} > 0.\end{aligned}$$

For $£250,000 < p <= £500,000$, I have $b_1 >= p > 250,000$, then:

$$\begin{aligned}\Delta S_{p2} &= (1-\alpha_1)b_1 - (1-\alpha_0)\frac{b_0+10000}{1+0.05} \\ &= \frac{[(1-\alpha_1)b_1 - (1-\alpha_0)b_0] + [(1-\alpha_1)0.05b_1 - (1-\alpha_0)10000]}{1.05} > 0.\end{aligned}$$

For $p > £500,000$: $\Delta S_{p2} = (1-\alpha_1)\frac{b_1+K_1}{1+\tau_1} - (1-\alpha_0)\frac{b_0+K_1}{1+\tau_1} > 0$.

Therefore, ΔS^S is positive in all price ranges. □

Chapter 5

Conclusions

With the emergence of big data and the UK government's commitment to open data, as well as copyright exceptions for non-commercial research, researchers now have the opportunity to gather data from various sources, merge them, and utilise them to address formerly challenging research problems within a single individual's workload. This thesis capitalises on these developments to construct several innovative datasets and offer fresh insights into the UK housing market. It demonstrates a research framework that is less constrained by data compared to conventional studies in the field. The approach enables the incorporation of information from new sources into the analysis to tackle issues encountered during the research process.

In this chapter, I present a concise summary of the research questions and principal contributions of each chapter, as well as suggest potential extensions worth investigating. The focus of Chapter 2 is on the novel framework for constructing extensive property-level micro-datasets employing big data techniques. It emphasises the four primary sources, highlighting crucial variables or information in each source. For instance, the CTB from council tax data plays a pivotal role in the analysis in Chapter 3 as one of the instruments for overcoming the simultaneity between TOM and Price. It also serves as a fixed-effect control in Chapter 4 to reduce the standard errors in estimations. This is because the CTB is determined by the VOA using consistent criteria across the country, rendering it a reliable indicator of property value and ensuring it is not influenced by agents involved in transactions such as buyers, sellers, and agents. The variables generated from linking records from various sources are also essential to this study. For example, TOM is calculated as the difference between the transfer date obtained from PPD data and the first published date acquired from Zoopla listing data. The overpricing proxy employed in Chapter 3 is derived in a similar manner. Without these variables, it would be impossible to identify the Price-TOM relationship.

Chapter 2 subsequently delineates the data integration method I specifically devised for this research, which encompasses a combination of text matching and unique identifier matching algorithms. Utilising these algorithms, I generated multiple datasets for the investigations in Chapters 3 and 4. The smallest dataset incorporates information on one-third of the population's residential transactions, while the largest dataset encompasses over 92% of all transactions documented by the Land Registry during my sample period. To the best of my knowledge, these datasets are among the most comprehensive employed in similar studies.

In Chapter 3, I re-examine the relationship between price and TOM in the housing market, which has been a long-standing puzzle in the field. Housing plays a significant role in household portfolios, but its heterogeneity and illiquidity make it challenging for market participants to determine the true value of a property. The Hedonic model suggests that properties are valued for their utility-bearing attributes such as physical characteristics and location-related amenities and services (Rosen 1974). However, even after controlling for these attributes, prices remain dispersed rather than uniform in the local market (He et al. 2017).

The search and matching theory of housing markets formalises a framework for understanding the process of searching for a property's true value and the resulting equilibrium market price. According to this theory (Anglin et al. 2003; Krainer and LeRoy 2002; Wheaton 1990), the price and TOM depend simultaneously on the probability of sale, indicating a positive relationship between them (Hayunga and Pace 2019). However, the empirical evidence for this relationship is less clear. The impact of TOM on the price and the impact of the price on TOM have been extensively researched over the past three decades with inconclusive results prior to

2015 (Benefield et al. 2014; Johnson et al. 2007). Despite several studies (Dubé and Legros 2016; Hayunga and Pace 2019; He et al. 2017) conducted in recent years that have directly investigated this puzzle, a consensus has yet to be reached.

Inspired by the literature, I identify two primary sources of inconsistency in prior empirical studies of the relationship between price and TOM. Firstly, endogeneity arising from the joint determination of price and TOM has resulted in model identification issues. Secondly, the absence of a variable measuring overpricing, often omitted in previous studies due to data constraints, has also contributed to inconsistent results. By capitalising on my big data, I address the first issue by proposing a 2SLS estimation process with two novel instrumental variables and the second issue by including a newly constructed measure of overpricing in the model. I confirm a positive relationship between price and TOM through the simultaneous equation model, which aligns with the search theory in the housing market.

An unexpected finding emerges from the 2SLS results, demonstrating that properties listed as ”chain-free” on average sell for 4-5% less than ”in-chain” properties, all else being equal. However, compared to ”in-chain” sales, ”chain-free” is often considered a selling point in practice as it offers a more flexible and efficient buying process. Through an analysis of the mediation effect of the initial listing price, it has been observed that sellers who are free from chains—that is, they are not obligated to sell their current property before purchasing a new one—tend to set lower initial asking prices and obtain lower transaction prices. I attribute this to agency costs, which constitute the main factor affecting the setting of the initial listing price for properties not linked to a chain. Previous research by Levitt and Syverson (2008) indicates that agents tend to recommend a lower initial listing price for a quicker sale. Conversely, sellers who are part of a chain are often financially constrained, and their sale speed is dependent on the progression of other sales in the chain. Consequently, agents are less likely to persuade these sellers to set a lower initial list price due to their higher risk aversion, and quick sales are unlikely to occur due to the chain's nature. As a result, ”in-chain” sellers tend to have higher initial listing prices and experience fewer principal-agent problems compared to ”chain-free” sellers.

Two potential avenues for future exploration are proposed to build upon the aforementioned findings. Firstly, new algorithms could be developed to identify unsold listed properties, which can then be combined with existing datasets to perform survival analysis, thereby directly investigating the relationship between the probability of sale, transaction price, and TOM in search theory. Secondly, the popularity of online estate agents, which enable homeowners to advertise their properties on property portals like Rightmove or Zoopla for a fixed fee, is increasing among UK home-sellers. It would be intriguing to examine whether these online estate agents reduce information asymmetry in the housing market.

In Chapter 4, I examine the effects of property transaction tax changes, specifically the 2020 stamp duty land tax holiday, on the housing market in the UK. Property transaction taxes are common in many countries and are levied on the purchase of real estate. Stamp duty in the UK serves both as a source of government revenue and as a tool for regulating the housing market. The proceeds generated from stamp duty finance public programmes and services, including education, healthcare, and infrastructure development. Moreover, it helps prevent housing bubbles by making it challenging for speculators to engage in excessive buying and selling, thereby promoting long-term ownership and fostering stable communities. By requiring property owners to contribute to the public good, rather than solely for personal gain, stamp duty serves to balance the interests of individuals with the needs of the wider society.

However, stamp duty can also have negative effects by reducing the expected benefits of transactions for both buyers and sellers. It discourages trades, making it more difficult for properties to be held by those who value them most. Previous research has criticised transaction taxes for their adverse impact on mobility, hindering individuals from relocating for better opportunities and leading to negative consequences on employment and productivity (Hilber and Lyytikäinen 2017; Van Ommeren and Van Leuvensteijn 2005). Furthermore, the frequency of property transactions varies significantly across regions and households, but there is no compelling reason for imposing excessive taxes on frequently traded residential properties (Adam 2011).

In light of these criticisms and to support homeownership, the UK government reformed the stamp duty in December 2014, transitioning from a ”slab” system to a ”slice” system, which ultimately led to a decrease in taxes for the majority of taxpayers. However, despite numerous studies on the ”slab” stamp duty, the effects of the ”slice” system remain relatively unexplored. This study aims to address this gap by thoroughly examining the effects of the new tiered, progressive tax system on the residential market.

In 2020, the UK government introduced a temporary reduction in stamp duty as part of its job creation measures in response to the slowdown in housing activities due to the early outbreak of COVID-19 and the first national lockdown. In this study, I use this policy as a quasi-natural experimental setting to evaluate the tax implications on prices, trading patterns, and liquidity in the housing market.

In this chapter, I propose a Nash bargaining model to explain the ”slice” stamp duty system and demonstrate that a tax holiday can result in higher prices and increased surplus for sellers who trade during the tax holiday. I then estimate a series of DiD models and find that, on average, the SDH led to a 53% increase in housing transactions, a 60% rise in listing prices, and an over 2% increase in transaction prices. Additionally, I observe that sellers had stronger bargaining power as the SDLT holiday deadline approached. The entire tax savings from the SDH were passed on to sellers in the form of increased prices, reducing affordability for first-time buyers and home movers replacing their main residence. The results also provide evidence that market participants used the SDH to relocate away from highly urbanised polycentric areas during the Covid-19 pandemic. My findings show that while an SDH can stimulate market activity during an economic downturn and enable the housing market to adjust to changing conditions quickly, it also has the unintended consequence of reducing housing affordability.

As mentioned in Chapter 4, it is worth investigating whether the 2020 SDH is associated with the re-timing of transactions that would have occurred anyway or provided an additional boost (see Figure 4.6) in future studies. This aspect has not been explored in this study due to sample duration limitations.

Bibliography

Adam, Stuart (2011). *Tax by design: The Mirrlees review.* Vol. 2. Oxford University Press.

Akerlof, George A (1970). “The Market for” Lemons”: Quality Uncertainty and the Market Mechanism”. In: *The Quarterly Journal of Economics*, pp. 488–500.

Alexander, Carol and Michael Barrow (1994). “Seasonality and cointegration of regional house prices in the UK”. In: *Urban Studies* 31.10, pp. 1667–1689. DOI: 10.1080/00420989420081571.

Allen, Marcus T., Sheri Faircloth, and Ronald C. Rutherford (2005). “The Impact of Range Pricing on Marketing Time and Transaction Price: A Better Mousetrap for the Existing Home Market?” In: *The Journal of Real Estate Finance and Economics* 31.1, pp. 71–82. DOI: 10.1007/s11146-005-0994-4.

Anglin, Paul M., Ronald Rutherford, and Thomas M. Springer (2003). “The Trade-off between the Selling Price of Residential Properties and Time-on-the-Market: The Impact of Price Setting”. In: *Journal of Real Estate Finance and Economics* 26.1, pp. 95–111. DOI: 10.1023/A:1021526332732.

Angrist, Joshua, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu (May 2017). “Economic Research Evolves: Fields and Styles”. In: *American Economic Review* 107.5, pp. 293–97. DOI: 10.1257/aer.p20171117.

Athey, Susan and Guido Imbens (2016). “Recursive partitioning for heterogeneous causal effects”. In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Athey, Susan and Guido W Imbens (2019). “Machine learning methods that economists should know about”. In: *Annual Review of Economics* 11, pp. 685–725.

Aydin, Erdal, Santiago Bohórquez Correa, and Dirk Brounen (2019). "Energy performance certification and time on the market". In: *Journal of Environmental Economics and Management* 98, p. 102270. DOI: `10.1016/j.jeem.2019.102270`.

Ball, Michael J (1973). "Recent empirical work on the determinants of relative house prices". In: *Urban studies* 10.2, pp. 213–233. DOI: `10.1080/00420987320080311`.

Benefield, Justin D., Christopher L. Cain, and Ken H. Johnson (2014). "A Review of Literature Utilizing Simultaneous Modeling Techniques for Property Price and Time-on-Market". In: *Journal of Real Estate Literature* 22.2, pp. 149–175. DOI: `10.5555/reli.22.2.k72g87h737x82068`.

Besley, Timothy, Neil Meads, and Paolo Surico (2014). "The incidence of transaction taxes: Evidence from a stamp duty holiday". In: *Journal of Public Economics* 119, pp. 61–70. DOI: `10.1016/j.jpubeco.2014.07.005`.

Best, Michael Carlos and Henrik Jacobsen Kleven (2017). "Housing market responses to transaction taxes: Evidence from notches and stimulus in the UK". In: *The Review of Economic Studies* 85.1, pp. 157–193. DOI: `10.1093/restud/rdx032`.

Brueckner, Jan, Matthew E Kahn, and Gary C Lin (2021). *A New Spatial Hedonic Equilibrium in the Emerging Work-from-Home Economy?* Working Paper 28526. National Bureau of Economic Research. DOI: `10.3386/w28526`.

Bucchianeri, Grace W. and Julia A. Minson (2013). "A homeowner's dilemma: Anchoring in residential real estate transactions". In: *Journal of Economic Behavior & Organization* 89, pp. 76–92. DOI: `10.1016/j.jebo.2013.01.010`.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru (2020). "Why is intermediating houses so difficult? evidence from ibuyers". In: *NBER Working Paper No. w28252*. DOI: `10.3386/w28252`.

Bun, Maurice J.G. and Frank Windmeijer (2011). "A comparison of bias approximations for the two-stage least squares (2SLS) estimator". In: *Economics Letters* 113.1, pp. 76–79. ISSN: 0165-1765. DOI: `https://doi.org/10.1016/j.econlet.2011.05.047`.

Campbell, John Y and Joao F Cocco (2007). "How do house prices affect consumption? Evidence from micro data". In: *Journal of monetary Economics* 54.3, pp. 591–621. DOI: `10.1016/j.jmoneco.2005.10.016`.

Cardella, Eric and Michael J. Seiler (2016). "The effect of listing price strategy on real estate negotiations: An experimental study". In: *Journal of Economic Psychology* 52, pp. 71–90. DOI: `10.1016/j.joep.2015.11.001`.

Chi, Bin, Adam Dennett, Thomas Oléron-Evans, and Robin Morphet (2021). "A new attribute-linked residential property price dataset for England and Wales, 2011 to 2019". In: *UCL Open: Environment Preprint*. DOI: `10.14324/111.444/000064.v1`.

Ciarlone, Alessio (2015). "House price cycles in emerging economies". In: *Studies in Economics and Finance*. DOI: `10.1108/SEF-11-2013-0170`.

Cook, Steven (2003). "The convergence of regional house prices in the UK". In: *Urban studies* 40.11, pp. 2285–2294. DOI: `10.1080/0042098032000123295`.

Cooper, Crispin, Scott Orford, Chris Webster, and Christopher B Jones (2013). "Exploring the ripple effect and spatial volatility in house prices in England and Wales: regressing interaction domain cross-correlations against reactive statistics". In: *Environment and Planning B: Planning and Design* 40.5, pp. 763–782. DOI: `10.1068/b37062`.

Dachis, Ben, Gilles Duranton, and Matthew A Turner (2012). "The effects of land transfer taxes on real estate markets: evidence from a natural experiment in Toronto". In: *Journal of economic Geography* 12.2, pp. 327–354. DOI: `10.1093/jeg/lbr007`.

Davidoff, Ian and Andrew Leigh (2013). "How do stamp duties affect the housing market?" In: *Economic Record* 89.286, pp. 396–410. DOI: `10.1111/1475-4932.12056`.

Delventhal, Matthew J., Eunjee Kwon, and Andrii Parkhomenko (2022). "JUE Insight: How do cities change when we work from home?" In: *Journal of Urban Economics* 127. JUE Insights: COVID-19 and Cities, p. 103331. DOI: `10.1016/j.jue.2021.103331`.

Dubé, Jean and Diègo Legros (2016). "A Spatiotemporal Solution for the Simultaneous Sale Price and Time-on-the-Market Problem". In: *Real Estate Economics* 44.4, pp. 846–877. DOI: 10.1111/1540-6229.12121.

Eerola, Essi, Oskari Harjunen, Teemu Lyytikäinen, and Tuukka Saarimaa (2021). "Revisiting the Effects of Housing Transfer Taxes". In: *Journal of Urban Economics*, p. 103367. DOI: 10.1016/j.jue.2021.103367.

Fritzsche, Carolin and Lars Vandrei (2019). "The German real estate transfer tax: Evidence for single-family home transactions". In: *Regional Science and Urban Economics* 74, pp. 131–143. DOI: 10.1016/j.regsciurbeco.2018.08.005.

Genesove, David and Christopher Mayer (2001). "Loss Aversion and Seller Behavior: Evidence from the Housing Market". In: *The Quarterly Journal of Economics* 116.4, pp. 1233–1260. DOI: 10.1162/003355301753265561.

Glower, Michel, Donald R. Haurin, and Patric H. Hendershott (1998). "Selling Time and Selling Price: The Influence of Seller Motivation". In: *Real Estate Economics* 26.4, pp. 719–740. DOI: 10.1111/1540-6229.00763.

Gray, David (2012). "District house price movements in England and Wales 1997–2007: An exploratory spatial data analysis approach". In: *Urban Studies* 49.7, pp. 1411–1434. DOI: 10.1177/0042098011417020.

Griffin, Beth Ann, Daniel F McCaffrey, Daniel Almirall, Lane F Burgette, and Claude Messan Setodji (2017). "Chasing balance and other recommendations for improving nonparametric propensity score models". In: *Journal of causal inference* 5.2, p. 20150026.

Hamermesh, Daniel S. (Mar. 2013). "Six Decades of Top Economics Publishing: Who and How?" In: *Journal of Economic Literature* 51.1, pp. 162–72. DOI: 10.1257/jel.51.1.162.

Han, Lu and Kevin D Sheedy (2022). *To Own Or to Rent?: The Effects of Transaction Taxes on Housing Markets.*

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer.

Hayunga, Darren K. and R. Kelley Pace (2019). "The Impact of TOM on Prices in the US Housing Market". In: *The Journal of Real Estate Finance and Economics* 58.3, pp. 335–365. DOI: 10.1007/s11146-018-9657-0.

He, Xin, Zhenguo Lin, Yingchun Liu, and Michael J. Seiler (2017). "Search Benefit in Housing Markets: An Inverted U-Shaped Price and TOM Relation". In: *Real Estate Economics*, pp. 1–36. DOI: 10.1111/1540-6229.12221.

Hilber, Christian AL and Teemu Lyytikäinen (2017). "Transfer taxes and household mobility: distortion on the housing or labor market?" In: *Journal of Urban Economics* 101, pp. 57–73. DOI: 10.1016/j.jue.2017.06.002.

Horowitz, Joel L. (1992). "The role of the list price in housing markets: Theory and an econometric model". In: *Journal of Applied Econometrics* 7.2, pp. 115–129. DOI: 10.1002/jae.3950070202.

Huang, Yunlong and Stanimira Milcheva (2020). "The Price–Time-on-Market Puzzle Revisited: Evidence from Big Data". In: *SSRN Working Paper*. DOI: 10.2139/ssrn.3837325.

Hudson, Chris, John Hudson, and Bruce Morley (2018). "Differing house price linkages across UK regions: A multi-dimensional recursive ripple model". In: *Urban Studies* 55.8, pp. 1636–1654. DOI: 10.1177/0042098017700804.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Johnson, Ken, Justin Benefield, and Jonathan Wiley (2007). "The Probability of Sale for Residential Real Estate". In: *Journal of Housing Research* 16.2, pp. 131–142. DOI: 10.1080/10835547.2007.12091978.

Jonathan, Halket, Mysliwski Mateusz, Nesheim Lars, and Simpson Polly (2018). "Estimating the benefits of transport investment". In.

Knight, John R. (2002). "Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes". In: *Real Estate Economics* 30.2, pp. 213–237. DOI: 10.1111/1540-6229.00038.

Kopczuk, Wojciech and David Munroe (2015). “Mansion tax: The effect of transfer taxes on the residential real estate market”. In: *American economic Journal: economic policy* 7.2, pp. 214–57. DOI: 10.1257/pol.20130361.

Krainer, John and Stephen F. LeRoy (2002). “Equilibrium valuation of illiquid assets”. In: *Economic Theory* 19.2, pp. 223–242. DOI: 10.1007/PL00004214.

Lai, Hang and Stanimira Milcheva (2021). “Long-run Discount Rates: Evidence from UK Repeat Sales Housing”. In: *SSRN Electronic Journal*, pp. 1–48. DOI: 10.2139/ssrn.3980392.

Lazear, Edward (1986). *Retail Pricing and Clearance Sales*. Tech. rep. 1. Cambridge, MA, pp. 14–32. DOI: 10.3386/w1446.

Levitt, Steven D. and Chad Syverson (2008). “Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions”. In: *Review of Economics and Statistics* 90.4, pp. 599–611. DOI: 10.1162/rest.90.4.599.

Liu, Xiaolong and Arno J. van der Vlist (2019). “Listing strategies and housing busts: Cutting loss or cutting list price?” In: *Journal of Housing Economics* 43.September 2018, pp. 102–117. DOI: 10.1016/j.jhe.2018.09.006.

McAvinchey, Ian D and Duncan Maclennan (1982). “A regional comparison of house price inflation rates in Britain, 1967-76”. In: *Urban Studies* 19.1, pp. 43–57. DOI: 10.1080/00420988220080041.

McGreal, Stanley, Paloma Taltavull de La Paz, Valerie Kupke, Peter Rossini, and Paul Kershaw (2016). “Measuring the influence of space and time effects on time on the market”. In: *Urban Studies* 53.13, pp. 2867–2884. DOI: 10.1177/0042098015596923.

Orford, Scott (2010). “Towards a data-rich infrastructure for housing-market research: deriving floor-area estimates for individual properties from secondary data sources”. In: *Environment and Planning B: Planning and Design* 37.2, pp. 248–264. DOI: 10.1068/b35082.

Parkinson, Aidan, Robert De Jong, Alison Cooke, and Peter Guthrie (2013). "Energy performance certification as a signal of workplace quality". In: *Energy Policy* 62, pp. 1493–1505. DOI: 10.1016/j.enpol.2013.07.043.

Petkova, Kunka and Alfons J Weichenrieder (2017). "Price and quantity effects of the German real estate transfer tax". In: DOI: 10.2139/ssrn.3004387.

Powell-Smith, A (2017). *House prices by square metre in England & Wales.*

Rosen, Sherwin (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". In: *Journal of Political Economy* 82.1, pp. 34–55. DOI: 10.1086/260169.

Scanlon, Kath, Christine Whitehead, and Fanny Blanc (2017). "A taxing question: Is Stamp Duty Land Tax suffocating the English housing market". In: *Report for Family Building Society.*

Slemrod, Joel, Caroline Weber, and Hui Shan (2017). "The behavioral response to housing transfer taxes: Evidence from a notched change in DC policy". In: *Journal of Urban Economics* 100, pp. 137–153. DOI: 10.1016/j.jue.2017.05.005.

Szumilo, Nikodem and Franz Fuerst (2015). "Who captures the "green value" in the US office market?" In: *Journal of Sustainable Finance & Investment* 5.1-2, pp. 65–84. DOI: 10.1080/20430795.2015.1054336.

Taylor, Curtis R. (1999). "Time-on-the-market as a sign of quality". In: *Review of Economic Studies* 66.3, pp. 555–578. DOI: 10.1111/1467-937X.00098.

Van Ommeren, Jos and Michiel Van Leuvensteijn (2005). "New evidence of the effect of transaction costs on residential mobility". In: *Journal of regional Science* 45.4, pp. 681–702. DOI: 10.1111/j.0022-4146.2005.00389.x.

Wheaton, William C (1990). "Vacancy, Search, and Prices in a Housing Market Matching Model". In: *Journal of Political Economy* 98.6, pp. 1270–1292. DOI: 10.1086/261734.

Wit, Erik R. de and Bas van der Klaauw (2013). "Asymmetric information and list-price reductions in the housing market". In: *Regional Science and Urban Economics* 43.3, pp. 507–520. DOI: 10.1016/j.regsciurbeco.2013.03.001.

Yavas, Abdullah and Shiawee Yang (1995). “The Strategic Role of Listing Price in Marketing Real Estate: Theory and Evidence”. In: *Real Estate Economics* 23.3, pp. 347–368. DOI: 10.1111/1540-6229.00668.

Young, Alwyn (2022). “Consistency without inference: Instrumental variables in practical application”. In: *European Economic Review* 147, p. 104112. DOI: 10.1016/j.euroecorev.2022.104112.

Appendix

A.Summary of Data by Different Splits

Table 5.1: Summary Statistics group by Rural-Urban Classification

RUC		Rural			Urban City&Town			Urban Conurbation	
Variable	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Transaction price	320691.6	275000	191512.6	282276.6	247000	167248.3	358149.6	278000	280860.4
Initial listing price	339202.3	290000	206155.4	296024.9	259950	177391.8	376748.8	290000	303533.4
Price spread	10950.5	5000	25062.8	7803	5000	18476.6	11011.3	5000	32625
TOM	241.2	183	180.6	216.3	168	158.2	217.8	169	158.8
Total floor area	109.3	97	48.6	95.7	87	37.9	96.9	88	38.2
Final listing price	331642.2	284000	201080.1	290079.7	250000	173414.2	369160.9	285000	295680.3
Num. habitable rooms	5.2	5	1.6	4.8	5	1.5	4.8	5	1.5
Num. open fireplaces	0.2	0	0.5	0.1	0	0.4	0.1	0	0.5
Current energy efficiency	61.4	63	14.2	63.8	65	11.2	62.4	64	11.2
Potential energy efficiency	80.7	82	9.7	81	82	7.4	80.2	82	7.4
Environment impact current	57.7	59	15.8	60.1	61	13.2	58.3	59	13
Environment impact potential	77.1	79	11.7	78.5	80	9.3	77.6	79	9.3
Energy consumption current	254.5	233	123.5	253.1	239	101.8	261.4	250	95.7
Energy consumption potential	123.2	110	80.1	126	112	68.5	133.4	121	65.2
Categorical Variable	Count	Percentage		Count	Percentage		Count	Percentage	
Property type	162911			419478			232548		
... House	157528	97%		386724	92%		202046	87%	
... Flat	5383	3%		32754	8%		30502	13%	
Duration	162911			419478			232548		
... freehold	153780	94%		373411	89%		176211	76%	
... leasehold	9131	6%		46067	11%		56337	24%	
New built	162911			419478			232548		
... no	162415	100%		419029	100%		232357	100%	
... yes	496	0%		449	0%		191	0%	
Price modifier	162911			419478			232548		
... fixed price	723	0%		1422	0%		688	0%	
... guide price	122947	75%		311224	74%		165519	71%	
... offers around	11419	7%		27395	7%		21728	9%	
... offers over	27594	17%		78988	19%		44401	19%	
... price on request	228	0%		449	0%		212	0%	
Chainfree	162911			419478			232548		
... FALSE	117183	72%		286376	68%		148506	64%	
... TRUE	45728	28%		133102	32%		84042	36%	
Garage	162911			419478			232548		
... FALSE	68903	42%		213476	51%		145223	62%	
... TRUE	94008	58%		206002	49%		87325	38%	
Driveway	162911			419478			232548		
... FALSE	82937	51%		241592	58%		148817	64%	
... TRUE	79974	49%		177886	42%		83731	36%	
Garden	162911			419478			232548		
... FALSE	12291	8%		46905	11%		35978	15%	
... TRUE	150620	92%		372573	89%		196570	85%	

Table 5.2: Summary Statistics group by Property Type

Property type		House			Flat	
Numerical Variable	Mean	Median	SD	Mean	Median	SD
Transaction price	315507.2	262000	213111.3	269200.4	209000	205609.6
Initial listing price	331291.5	275000	228218.4	288549.5	220000	227934.9
Price spread	9160.5	5000	23968.7	11383.5	5000	31023.6
TOM	218.5	169	161.2	256.6	200	181.9
Total floor area	101.9	91	40.6	64.6	61	22.1
Final listing price	324667.7	270000	222826.1	280583.9	215000	219749.2
Num. habitable rooms	5.1	5	1.4	2.9	3	0.8
Num. open fireplaces	0.2	0	0.5	0.1	0	0.3
Current energy efficiency	62.4	64	11.8	68.8	71	11.4
Potential energy efficiency	81.1	82	7.9	76.6	78	6.7
Environment impact current	58.4	59	13.5	66.5	68	13.9
Environment impact potential	78.3	80	9.8	74.3	77	10
Energy consumption current	256.5	242	103.1	247.1	224	122.4
Energy consumption potential	122	111	65.6	187.8	164	88.1
Categorical Variable	Count	Percentage		Count	Percentage	
RUC	746298			68639		
... Rural	157528	21%		5383	8%	
... Urban City&Town	386724	52%		32754	48%	
... Urban Conurbation	202046	27%		30502	44%	
Duration	746298			68639		
... freehold	701936	94%		1466	2%	
... leasehold	44362	6%		67173	98%	
New built	746298			68639		
... no	745419	100%		68382	100%	
... yes	879	0%		257	0%	
Price modifier	746298			68639		
... fixed price	2514	0%		319	0%	
... guide price	545683	73%		54007	79%	
... offers around	57438	8%		3104	5%	
... offers over	139821	19%		11162	16%	
... price on request	842	0%		47	0%	
Chainfree	746298			68639		
... FALSE	512769	69%		39296	57%	
... TRUE	233529	31%		29343	43%	
Garage	746298			68639		
... FALSE	367993	49%		59609	87%	
... TRUE	378305	51%		9030	13%	
Driveway	746298			68639		
... FALSE	406602	54%		66744	97%	
... TRUE	339696	46%		1895	3%	
Garden	746298			68639		
... FALSE	60303	8%		34871	51%	
... TRUE	685995	92%		33768	49%	

Table 5.3: Summary Statistics group by Transfer Before or During the SDH

Transaction during SDH		No			Yes	
Variable	Mean	Median	SD	Mean	Median	SD
Transaction price	297334.5	246000	203255.7	344260.5	288000	230101.1
Initial listing price	313769.4	259950	219750.6	359543.6	300000	244400.9
Price spread	9477.7	5000	23382.2	9050.5	5000	27323.9
TOM	204.3	164	137.1	261.7	189	205.9
Total floor area	97.2	88	39.4	102.4	91.3	43.4
Final listing price	306812.2	250000	213691.5	353311	295000	239552.6
Num. habitable rooms	4.8	5	1.5	5	5	1.5
Num. open fireplaces	0.1	0	0.5	0.1	0	0.4
Current energy efficiency	62.6	64	12	63.6	65	11.7
Potential energy efficiency	80.4	82	8.1	81.4	82	7.3
Environment impact current	58.8	60	13.8	59.8	61	13.6
Environment impact potential	77.6	79	10.1	78.8	80	9.3
Energy consumption current	259.9	245	106.4	246.3	234	100.9
Energy consumption potential	131.4	116	73.1	118.9	110	62.2
Categorical Variable	Count	Percentage		Count	Percentage	
RUC	567076			247861		
... Rural	110723	20%		52188	21%	
... Urban City&Town	294918	52%		124560	50%	
... Urban Conurbation	161435	28%		71113	29%	
Property type	567076			247861		
... House	518176	91%		228122	92%	
... Flat	48900	9%		19739	8%	
Duration	567076			247861		
... freehold	488036	86%		215366	87%	
... leasehold	79040	14%		32495	13%	
New built	567076			247861		
... no	566257	100%		247544	100%	
... yes	819	0%		317	0%	
Price modifier	567076			247861		
... fixed price	2109	0%		724	0%	
... guide price	416144	73%		183546	74%	
... offers around	43066	8%		17476	7%	
... offers over	105134	19%		45849	18%	
... price on request	623	0%		266	0%	
Chainfree	567076			247861		
... FALSE	391418	69%		160647	65%	
... TRUE	175658	31%		87214	35%	
Garage	567076			247861		
... FALSE	300100	53%		127502	51%	
... TRUE	266976	47%		120359	49%	
Driveway	567076			247861		
... FALSE	332533	59%		140813	57%	
... TRUE	234543	41%		107048	43%	
Garden	567076			247861		
... FALSE	67899	12%		27275	11%	
... TRUE	499177	88%		220586	89%	

Table 5.4: Summary Statistics group by Listed Before or During the SDH

Listed during SDH		No			Yes	
Variable	Mean	Median	SD	Mean	Median	SD
Transaction price	305993.6	252500	209204.5	361799.2	305000	238496.4
Initial listing price	322024.9	265000	224795.8	370520.4	310000	247003.7
Price spread	9525.5	5000	24423.9	5177.9	3000	20138
TOM	218	171	152	138.2	136	48.2
Total floor area	98.2	88.2	40.2	102.7	92	42
Final listing price	315519.1	260000	219535.6	366977.1	305000	244040.9
Num. habitable rooms	4.9	5	1.5	5	5	1.5
Num. open fireplaces	0.1	0	0.4	0.1	0	0.5
Current energy efficiency	63	64	11.9	63.7	65	11.3
Potential energy efficiency	80.8	82	7.9	81.7	82	7
Environment impact current	59.1	60	13.7	59.8	61	13.3
Environment impact potential	78	80	9.8	79	80	9
Energy consumption current	255.9	242	104.7	244.6	233	97.6
Energy consumption potential	127.1	114	69.6	116.1	109	57.8
Categorical Variable	Count	Percentage		Count	Percentage	
RUC	583540			92161		
... Rural	115714	20%		18273	20%	
... Urban City&Town	300899	52%		46640	51%	
... Urban Conurbation	166927	29%		27248	30%	
Property type	583540			92161		
... House	535658	92%		86327	94%	
... Flat	47882	8%		5834	6%	
Duration	583540			92161		
... freehold	504516	86%		82048	89%	
... leasehold	79024	14%		10113	11%	
New built	583540			92161		
... no	582678	100%		92078	100%	
... yes	862	0%		83	0%	
Price modifier	583540			92161		
... fixed price	1893	0%		223	0%	
... guide price	429627	74%		69286	75%	
... offers around	43434	7%		6201	7%	
... offers over	107977	19%		16366	18%	
... price on request	609	0%		85	0%	
Chainfree	583540			92161		
... FALSE	382074	65%		61598	67%	
... TRUE	201466	35%		30563	33%	
Garage	583540			92161		
... FALSE	305605	52%		46683	51%	
... TRUE	277935	48%		45478	49%	
Driveway	583540			92161		
... FALSE	337159	58%		51035	55%	
... TRUE	246381	42%		41126	45%	
Garden	583540			92161		
... FALSE	66694	11%		9017	10%	
... TRUE	516846	89%		83144	90%	