Essays in Urban Economics

Margarida Madaleno

Department of Geography and Environment

London School of Economics

A thesis submitted to the Department of Geography and Environment of the London School of Economics for the Degree of Doctor of Philosophy

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 23,000 words.

Margarida Madaleno

Statement of joint work

I certify that Chapter 3 of this thesis is co-authored with Stephen Law. I contributed to 50 % of the work. Chapter 4 of this thesis is co-authored with Joel Han. I contributed to 50 % of the work.

Acknowledgements

My first thank you goes to my supervisors. Henry. Working at WWC was the best professional experience I've ever had (and not because I haven't had many professional experiences, as Max noted in 2015). Thanks to you I was able to do the PhD at LSE. It was often painful, as growth usually is, but your pastoral care and empathy helped me through. Steve. Whenever you come up in conversation I always refer to you as the Greatest of All Time (GOAT). I've learned so much from you, mainly because you had the patience to unpack my econometric takes (which were often "creative").

My second thank you goes to LSE Geography faculty. Olmo. You assured me that I had done rather well during EC400, even though I felt completely defeated - thank you for your support throughout. Neil. Teaching your course was so enjoyable, and I appreciate the opportunity you gave me to deliver a lecture on Europe's ageing population (where I made a half-joke about illegal care homes that elicited bewilderment more than laughter). Felipe. Thanks for bearing with me through Quant methods - I'm certain that, in a rare case, I genuinely learned more than the students.

My third thank you goes to my PhD friends, who made the journey worthwhile. Everyday I'd wake up with a sense of purpose and looked forward to our leisurely lunches. Furthermore, learning from the human geographers (as opposed to us, the robot geographers) was a unique experience that I will forever treasure. Louise, Yohann, Lars, Matt, Andreas, Lindsey, Paul, Jams, Tanner, Carl, Yoonai, Didi, Carwyn.

My fourth thank you goes to my family. Mom. You made every sacrifice so that I could get a great education. Dad. The past ten years were difficult but you were with me the whole way. Thank you to my siblings: Chico, Luis, Ze, Pedro, and Mariana. Thank you also to my grandparents and cousins. Finally, thank you to a friend who is like family - Chantal.

Abstract

This thesis consists of four policy impact evaluations that consider outcomes related to marriage, mental health, architectural valuation, and parenting. The first of these impact evaluations seeks to understand the extent to which accessing housing credit affects the decision to marry. To this end, it evaluates the impact of a series of maps, which the US government created in order to guide lending in the wake of the great depression. A federal government agency – known as the Home Owners Loan Corporation (HOLC) – assigned a grade to each neighbourhood block on a four point scale (A, B, C, D). These grades aimed to capture the extent to which houses in these blocks would appreciate or depreciate in the future. The maps were then circulated to banks. This chapter evaluates the impact of these maps on area-level marriage, finding that discriminated neighbourhoods had fewer married individuals. It also concludes that these effects are due to the mortgage mechanism per se, and are not simply a product of sorting. The second of these impact evaluations considers the effect of the covid-19 pandemic on mental health. It exploits quasi-exogenous variation in lockdown severity in the UK, paired with high-frequency mental health data derived from Tweets. It finds that the lockdowns did, indeed, have a negative effect on mental health. The third of these impact evaluations considers the extent to which architectural design is capitalised into commercial rents, with the goal of understanding whether passers-by value aesthetic features of the public realm. To this end, it uses Google Streetview data on the ornateness of London facades, and uses a town centre fixed effects specification. It finds that architectural ornateness of facades is capitalised into commercial rents, and that individuals are willing to pay for distinctive architecture within the public realm. The final impact evaluation considers the effect of neighbourhood crime on parenting, comparing effects for different populations. It finds that residents of social housing do not change parenting practices in response to increased crime, while private sector housing families do. These findings may be suggestive of dynamic complementarities between neighbourhoods and parenting investments.

Contents

1 Introduction

This thesis consists of four independent chapters on urban economics.

The first chapter considers the following reduced form question: does access to housing credit affect the decision to marry? To this end, it evaluates the impact of an area-based housing credit discrimination policy, popularly known as redlining, on area-based marriage. In the 1930s, the federal US government conducted block-level data collection in order to determine the likely trajectories of home prices across several large cities. This data collection culminated in the creation of "redlining" maps, which were covertly circulated to banks and guided lending. Ultimately, as previous research has shown, these maps catalysed neighborhood decline in disfavoured areas. This study shows that discrimination also led to a decline in marriage, specifically through housing credit mechanism. Furthermore, it concludes that these area-level effects are not simply a product of sorting.

The second chapter considers the impact of the covid-19 pandemic on mental health. Faced with a public health crisis during the covid-19 pandemic, the UK government pursued a series of "lockdown" policies which prevented virus contagion. However, the mental health effects of the social isolation inherent to these policies were widely, but speculatively, cited as a failure of these policies. This study deploys Twitter data, from which mental health states - depression, anxiety, and stress - are inferred through text analysis methods, in conjunction with quasi-experimental methods, in order to understand the mental health impacts of these policies. This study contributes to pending issues in Twitter data representativeness, and finds that, overall, the lockdown policies had an adverse impact on mental health.

The third chapter evaluates whether architecture and design features are capitalised into commercial rents. It aims to understand whether individuals are willing to pay for aesthetics of the public realm, and whether there is economic value in preserving good architecture. This study contributes both to the planning debate, which has been marked by a paucity of evidence regarding the value of design, and to understanding how the public realm is valued. To this end, it uses a novel data set of facade "ornateness" for London town centres. To circumvent the endogeneity of architecture, it uses town centre fixed effects and a geographic control function at the town centre level. It finds that ornateness is capitalised into commercial rents, particularly for streets that are non-ornate and use older materials.

The fourth chapter considers how parenting behavior responds to changes in the neighborhood environment, focusing on reconciling the disparate findings across two different data samples and estimation methods. In particular, it compares individuals in the Moving to Opportunity (MTO) experiment to individuals in the Project for Human Development in Chicago Neighborhoods (PCDHN) sample. It finds that neighborhoods affect parenting differently across these two samples, despite efforts to conduct a like-for-like comparison. These findings may point to potential dynamic complementarities between family and neighborhood investments, particularly with respect to families in social housing. However, this study does not isolate this mechanism, with effects being potentially driven by confounders that simultaneously affect the choice to opt-into social housing and the effects of neighborhoods on parenting.

2 Housing credit and marriage: evidence from redlining

2.1 Introduction

There is little evidence on whether obtaining housing credit, and purchasing a home, is an important determinant of marriage. As noted by Eriksen (2010), the classic Gary Becker framework for marriage decisions (Becker (1973)) is ambiguous about the effects of housing wealth on marriage. Two opposing channels may lead to these ambiguous effects. On one hand, there are so-called "independence" effects whereby home-ownership renders single households more productive, leading to a negative effect on marriage. On the other hand, economies of scale in household production and housing consumption may lead to an a positive impact of home-ownership on marriage - i.e. owning a home (rather than renting) makes the household more productive, and therefore, marriage more attractive. Furthermore, it is also possible that, in the marriage market, home-ownership (and the wealth that this implies) makes individuals more "marriageable".

Empirical evidence on the impact of home-ownership - and housing credit specifically - on marriage is relatively scarce. This is largely due to the lack of exogenous variation in housing credit. In other words, the same factors that determine housing credit disbursal and rates also affect marriage. For instance, an individual's low income may disqualify them from obtaining housing credit, and simultaneously, make them a less attractive marriage partner. This study therefore exploits an unusual policy, which catalysed quasi-exogeneous area-based housing credit discrimination.

Still, causal evidence seems to support the idea that home-ownership is conducive to marriage, and the relationship holds in China (Hu & Wang (2019)) and the US (Eriksen (2010)). In particular, Eriksen (2010) find that home-ownership subsidies increase the likelihood of marriage. Similarly, Ricks (2021) find that the VA Loan Program, which subsidised housing credit for veterans in the 1950s, increased marriage. Furthermore, Hu & Wang (2019) note that in China, this causal relationship is driven by the greater attractiveness of home-owners in the marriage market. Finally, Miller & Park (2018) consider whether marriage leads to more home-ownership, exploiting quasi-random legalisation of same-sex marriage. They find that marriage does indeed lead to higher rates of home-ownership.

As noted, this study exploits quasi-exogenous variation in housing credit terms. In particular, it evaluates an area-based housing credit discrimination policy, popularly known as redlining. The Home Owners Loan Corporation (HOLC), a federal Roosevelt-era agency, created colour-coded maps for over 239 major US cities, which described the probable economic trajectory of each neighbourhood block. These maps were then covertly circulated to banks, and ultimately guided lending. As prior redlining impact evaluations have shown, these maps changed the course of neighbourhood trajectories (Krimmel (2018) Aaronson et al. (2017)).

In terms of the methodology used to uncover the reduced form effects, which allow us to make inferences about housing credit and marriages, the policy is superficially amenable to a spatial regression discontinuity design (RDD) set-up. This is due to the fact that the maps represent sharp yet contiguous discontinuities in grades. However, it will be shown that these grades were not randomly assigned, and were in fact based on careful data collection. More crucially, in addition to being based on careful collection, the grade designations followed pre-existing discontinuities in the variables featured in HOLC's data surveys. This means that the basic RDD requisite of variable smoothness across the boundary does not hold.

To circumvent this, a spatial RDD-propensity score matching (PSM) approach is deployed. In particular, this entails adapting the PSM approach to a spatial set-up, as in Keele et al. (2015). Loosely, census tracts of differing grades that share a border are compared, whilst attributing more weight to border pairs that are comparable with respect to pre-treatment variables. This is achieved by matching exactly with respect to the border fixed effects, and then applying standard PSM approaches to match on pre-treatment covariates.

The study finds that the policy decreased marriage in discriminated areas. However, several mechanisms could be driving these results, and the main regressions do not isolate the housing credit mechanism. For instance, it is possible that the HOLC policies led to neighbourhood decline, which in turn decreased marriage. In this case, the area-level effects previously outlined would not be due to the housing credit mechanism. In order to understand whether alternative mechanisms explain the main results, this study presents the main regressions with and without a variable that controls for neighbourhood decline. In this case, neighbourhood decline is summarised by a variable that captures house prices. It is held that if the results remain statistically signficant in spite of the inclusion of the house price variable, then the results are explained by the housing credit mechanism per se, rather than neighbourhood decline. Indeed, the results are robust to the inclusion of this variable, which suggests that the main results are driven by housing credit per se.

In order to understand whether the results simply reflect sorting, a city-level analysis is performed. It is held that is possible that the housing credit discrimination simply caused never-married individuals to sort into discriminated areas. However, the city-level results show that cities that received the maps experience more marriage decline than cities that did not receive the maps. In sum, the results are not a feature of sorting, and instead suggest that poor access to housing credit indeed decreased marriage.

The study concludes with some implications.

2.2 Policy overview

This study is interested in understanding whether access to housing credit influences marriage decisions. However, as noted, housing credit access is endogenous to family outcomes. For instance, poverty may prevent someone from obtaining housing credit, while simultaneously make them less attractive marriage partners. This question has not been widely studied precisely due to this endogeneity problem. Answering it would require an exogenous source of housing credit discrimination, but to my knowledge, there are few, if any, experiments that have randomly allocated housing credit conditions. The HOLC maps therefore present a quasi-exogenous source of variation in housing credit that allows us to understand the causal impact of housing credit on marriages. This section outlines some policy detail.

As noted by Hillier (2005), the HOLC was created by the federal government to slow the inevitable home foreclosures caused by the Great Depression. This institution was initially conceived in order to provide more favourable housing credit and loans to struggling families, and it did so between 1933 and 1936. In 1935, the Federal Home Loan Bank Board (FHLBB) used HOLC staff to conduct surveys of the desirability of neighborhoods within 239 cities. The objective of this work was to understand the trajectories of different neighbourhoods in order to better understand the viability of the aforementioned housing credit payments. Throughout this City Survey Program, HOLC surveyors collected data and conducted qualitative analyses for each neighbourhood block. This data collection culminated in the drafting of maps for chosen cities. In particular, the maps outlined which parts of the cities were to be lent to and which parts of the city were considered dangerous to loan to. In practice, each block was assigned a grade from "A" (most desirable) to "D" (least desirable), and the maps are colour-coded to reflect this. Green designated "A" neighbourhoods, blue designated "B" neighbourhoods, yellow designated "C" neighbourhoods, and finally red designated "D" neighbourhoods. Although the maps were arguably made for internal consumption by HOLC, the maps were then covertly (in the sense that the general public did not know about them until much later on) circulated to lending institutions. Given that these maps were designed to predict the neighbourhood trajectories, it is not obvious that differences between neighbourhoods after the treatment period are due to the maps per se. However, recent causal evidence shows that the maps did indeed catalyse neighbourhood decline in "D" graded areas (Krimmel (2018), Aaronson et al. (2017)). Furthermore, in an analysis of housing credit disbursal in Philadelphia, Hillier (2003) shows that it was only in "D" graded areas that the maps were binding, in the sense that lenders offered more stringent housing credit terms than they would have otherwise.

This study considers years 1950, 1960, and 1970 as post-treatment years. The final post-treatment year is 1970, largely because the Fair Housing Act (FHA) of 1968 effectively rendered redlining illegal. However, as noted by Massey (2015), while the FHA curtailed discrimination, it did not fully end it. In practice, it was very difficult to legally contest housing discrimination, and those found guilty of it faced few repercussions. Still, it would be expected that the policy's impact reaches a maximum in 1970 and declines thereafter.

Figure 1 shows a HOLC map for Pittsburgh, Pennsylvania. It is clear from this map that "D" graded neighbourhoods predominantly border "C" (yellow) neighbourhoods. However, they also occasionally border "B" and "A" neighbourhoods as well.

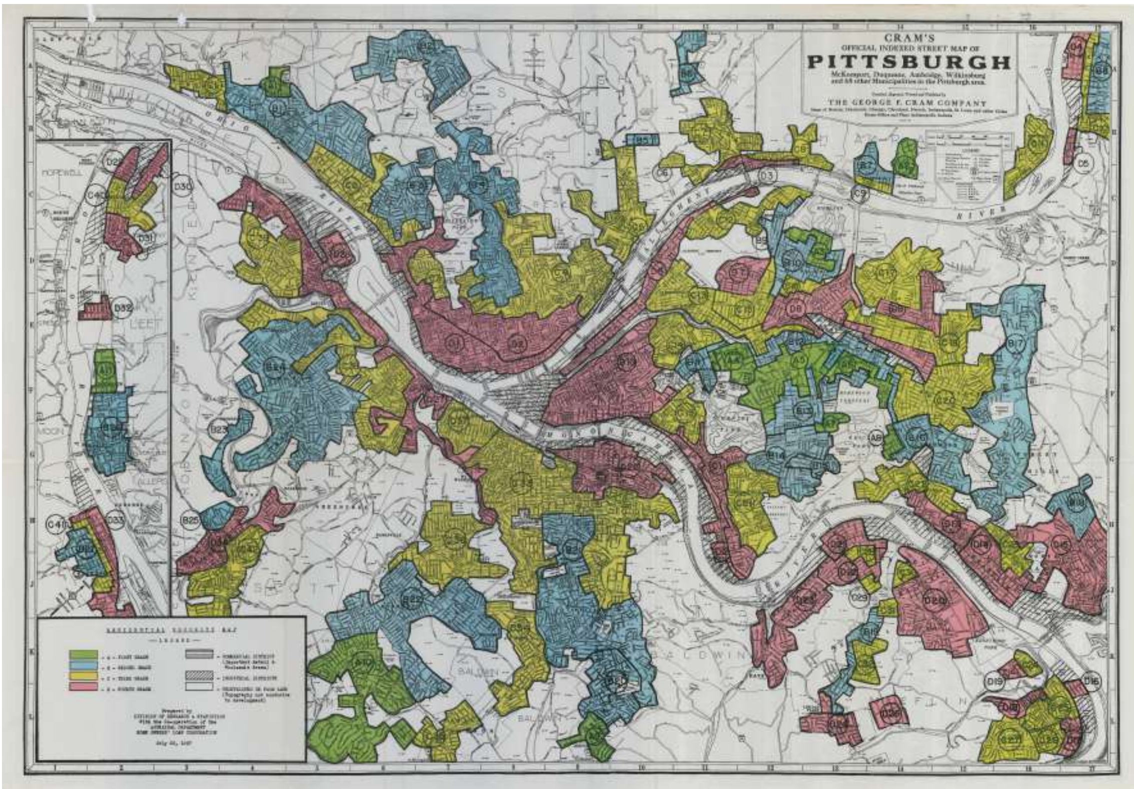


Figure 1: HOLC map for Pittsburgh, Pennsylvania

2.3 Data

This study exploits the Mapping Inequality project's digitization of the HOLC maps. In particular, this data includes block-level information on HOLC grades for 202 cities across the USA. Note that only 202 of 239 maps were digitised, due to unavailability of maps at the National Archives.

However, the outcome of interest, provided by Integrated Public Use Microdata Series' National Historical GIS (Manson et al. (2019)), is at the census tract level. In sum, the independent variable of interest (i.e. grade) was attributed at a finer spatial scale that the outcome (i.e. marriage). In order to bridge the grades at block-level with marriage at census tract level, the following steps were undertaken. Recall that the HOLC data was collected at the neighbourhood block level. Figure 2 presents a schematic diagram of the units of observations. In first instance, we have neighbourhood blocks as shown in Figure 2A. Each square of Figure 2A represents a neighbourhood block for which data was collected, with green representing a higher grade than blue. In second instance, contiguous sets of these same-graded blocks were dissolved into larger polygons. Figure 2B shows these two same-graded polygons. Finally, the polygons in Figure 2C were then intersected with the census tracts. These are our units of observation. Each number represents a different census tract which was overlayed with the polygons in Figure 2B to yield a unit of observation. Census tracts 1 to 3 represent unique, A-graded observations. Census tracts 4 to 7 represent unique B-graded observations. It is worth noting that since the blocks were dissolved into larger polygons, as shown in Figure 2C, a census tract can span several HOLC blocks. Census tract 8, on the other hand, spans both A and B grades. This census tract, and others like it, was dropped and is therefore represented by a line-patterned fill. In general, census tracts that spanned several grades were dropped, as the RDD set-up would yield effects that mechanically tended towards zero otherwise (i.e. differences across borders would be nonexistent due to same outcome variable).

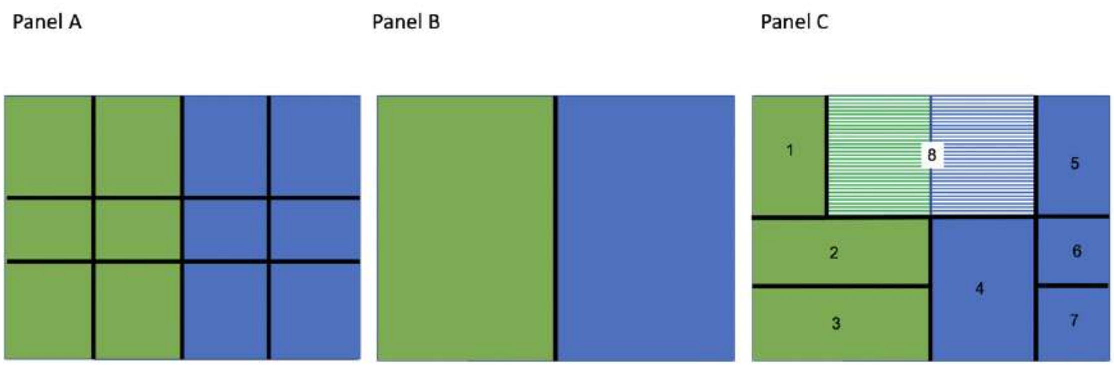


Figure 2: Schematic representation of units of observation

The main outcome variable is the percentage of married individuals at the census tract level. This statistic is calculated as the number of ever-married individuals (including married, divorced, and widowed), over the total number of individuals (same as the numerator but with single individuals). The dates that will be used in the post-treatment analysis include: 1950, 1960, and 1970. It is held that the maps were less binding after the Fair Housing Act of 1968, which rendered area-based housing credit discrimination illegal.

Table 1 presents the proportion of married individuals for each grade by year. Aligned with the progressive decline of marriage, for each grade, the marriage rate declines from 1950 to 1970. Other notable trends are that "D"-graded areas have less marriage that other grades. Furthermore, there seems to be a decrescendo in marriage as the areas become less desirable. Exceptions include 1960, where A and D were on a par, and 1950, where A was marginally worse than B. In general, it seems that marriage is positively correlated with area desirability.

Table 1: Descriptive statistics

	Marriage		
	1950	1960	1970
A	0.77	0.75	0.74
	(0.07)	(0.10)	(0.07)
Observations	445	438	1738
B	0.78	0.77	0.74
	(0.06)	(0.07)	(0.07)
Observations	1886	2000	6018
C	0.76	0.77	0.73
	(0.06)	(0.06)	(0.07)
Observations	6006	6268	16099
D	0.75	0.75	0.69
	(0.06)	(0.06)	(0.08)
Observations	8938	8949	17039

Note: Sets of three rows feature mean, standard deviation (in parentheses), and number of observations, for each grade.

2.4 Estimation

This section will describe the estimation strategy that will be deployed in order to understand the causal impacts of the HOLC maps on marriage. In particular, it will start by outlining the basic estimation strategy - the spatial RDD. It will discuss the estimating assumptions of this method, and how they are not satisfied. Subsequently, the PSM method is introduced as a way to overcome these estimation obstacles.

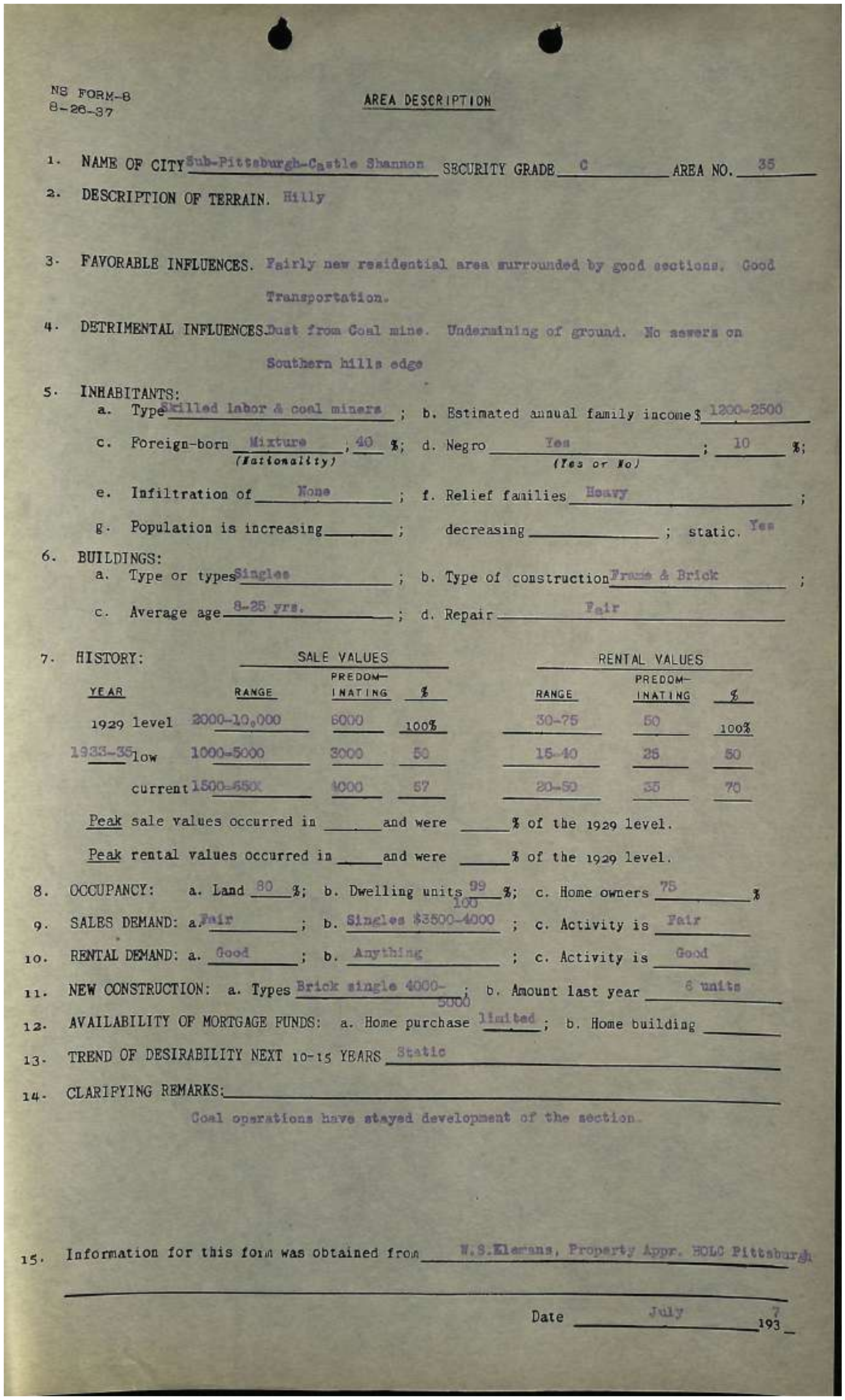
As is clear from Figure 1, there are sharp spatial discontinuities in treatment. Given this set-up, the natural estimation strategy would be to perform a spatial RDD, following Black (1999). As with any RDD, the basic intuition is that there is are adjacent observations that are similar in all respects except for their exposure to a policy. This policy is assigned according to a particular running variable, which in this case would be distance to a border. The policy effects are given by:

$$E[Y_1 - Y_2|X = c] = E[Y_1|X = c] - E[Y_2|X = c] = \lim_{x \downarrow c} E[Y|X = x] - \lim_{x \uparrow c} E[Y|X = x] \qquad (1)$$

Where Y_1 is the outcome for the treated group, Y_2 is the outcome for the untreated group, X is the running variable (i.e. distance to the D-A/D-B/D-C border), and c is the cut-off that marks the discontinuity in treatment (i.e. zero). Essentially the idea is to compare observations that approach the cut-off from either side.

2.4.1 RDD-PSM

Plans to perform a standard RDD are disrupted by the fact that these borders were not randomly drawn. In fact, the masses of colour in Figure 3 represent contiguous sets of neighbourhood blocks that were meticulously analysed by HOLC. For each block, surveyors would fill out a form that would depict its characteristics. Figure 3 presents one of these block-level surveys for Pittsburgh, Pennsylvania. This survey was collected in 1937, and the image was provided by the Mapping Inequality project. From this image, it is clear that HOLC was looking for specific things, and they were rather uniform in their data collection. The insights collected by the surveyors ranged from qualitative to quantitative. For instance, with respect to the nationality of foreign-born individuals, this particular block is described as a "mixture" where 40 % of individuals are non-native. Average family income at the block level is deemed between \$ 1200 and \$ 2500. There is a binary indicator for whether there are African-Americans in the block (here a "yes") as well as a percentage for the proportion of the population that is African-American (10 % in this case). And so on. The key point that is made in Figure 3 is that the borders were drawn far from randomly. This is not necessarily a problem, provided the borders are not drawn according to discontinuities in these variables. As will be shown, this is not the case. This ultimately justifies the PSM method that will be deployed in this paper, as measures need to be taken in order to circumvent the endogenously drawn borders.



NS FORM-8
8-26-37

AREA DESCRIPTION

1. NAME OF CITY Sub-Pittsburgh-Castle Shannon SECURITY GRADE C AREA NO. 35

2. DESCRIPTION OF TERRAIN. Hilly

3. FAVORABLE INFLUENCES. Fairly new residential area surrounded by good sections. Good Transportation.

4. DETRIMENTAL INFLUENCES Dust from Coal mine. Undermining of ground. No sewers on Southern hills edge

5. INHABITANTS:
a. Type Skilled labor & coal miners ; b. Estimated annual family income $ 1200-2500
c. Foreign-born Mixture (Nationality) , 40 %; d. Negro Yes (Yes or No) ; 10 %;
e. Infiltration of None ; f. Relief families Heavy ;
g. Population is increasing ____ ; decreasing ____ ; static. Yes

6. BUILDINGS:
a. Type or types Singles ; b. Type of construction Frame & Brick ;
c. Average age 8-25 yrs. ; d. Repair Fair

7. HISTORY:

YEAR	SALE VALUES RANGE	SALE VALUES PREDOMINATING	%	RENTAL VALUES RANGE	RENTAL VALUES PREDOMINATING	%
1929 level	2000-10,000	6000	100%	30-75	50	100%
1933-35 low	1000-5000	3000	50	15-40	25	50
current	1500-6500	4000	57	20-50	35	70

Peak sale values occurred in ____ and were ____ % of the 1929 level.
Peak rental values occurred in ____ and were ____ % of the 1929 level.

8. OCCUPANCY: a. Land 80 %; b. Dwelling units 99 100 %; c. Home owners 75 %

9. SALES DEMAND: a. Fair ; b. Singles $3500-4000 ; c. Activity is Fair

10. RENTAL DEMAND: a. Good ; b. Anything ; c. Activity is Good

11. NEW CONSTRUCTION: a. Types Brick single 4000-5000 ; b. Amount last year 6 units

12. AVAILABILITY OF MORTGAGE FUNDS: a. Home purchase limited ; b. Home building ____

13. TREND OF DESIRABILITY NEXT 10-15 YEARS Static

14. CLARIFYING REMARKS:
Coal operations have stayed development of the section.

15. Information for this form was obtained from W.S.Klerans, Property Appr. HOLC Pittsburgh

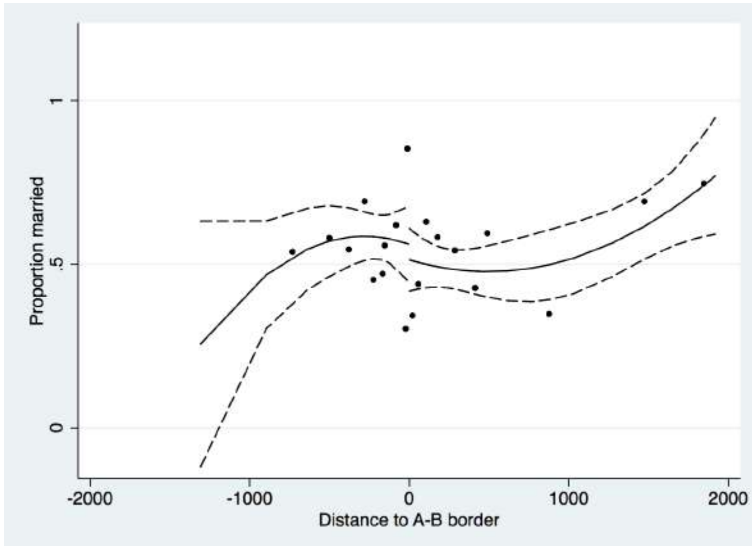
Date July 7 193_

Figure 3: Block-level survey for Pittsburgh, Pennsylvania

With respect to the standard RDD method, the identifying assumption here is that there are no discontinuities across the border for any variable except treatment. Formally this amounts to arguing that E[Y_1—X] and E[Y_2—X] are continuous at X=c. In practice this amounts to arguing that neighbours across the border are the same except for their exposure to policy. This would be the case if, for instance, the borders had been drawn randomly. As was discussed, the borders were based on careful block level surveys, which implies that this is likely not the case. Given the granularity of the surveys it is indeed unlikely.

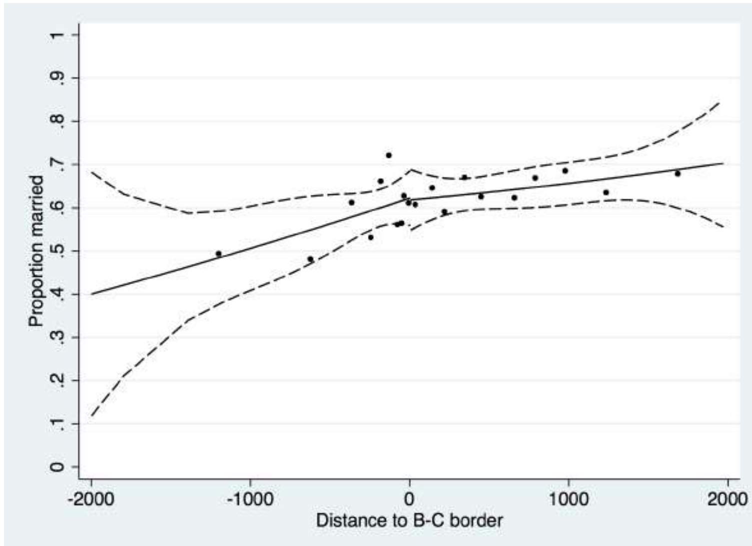
Figures 4 to 7 show pre-treatment (1930) graphs for differences across different border types. Indeed, there do not seem to be problematic discontinuities for A-B and B-C border pairs, but there are pre-existing discontinuities for C-D and ABC-D borders. Again, this is not surprising given that block grades were attributed on the basis of careful area-level data collection. It worth noting that there is a concern with respect to these graphs. This is due to the specific data set-up. In particular, recall that the observation is the census tract. Census tracts are defined as a given constant population, with differing physical area. This means that densely-populated places will have physically smaller census tracts. Recall also that distances to the border are calculated from the census tract centroid to the closest part of the border, meaning that physically smaller (i.e. more densely populated) census tracts will by construction be closer to the border. The graphs do not account for these data specificities, which means that in an extreme case, observations in the bins surrounding the border may belong to one densely populated city (e.g. New York), while bins farther away may correspond to a sparsely populated city. This is problematic if spatial trends differ across cities.

Figure 4: Pre-treatment marriage cross-border comparisons: A-B



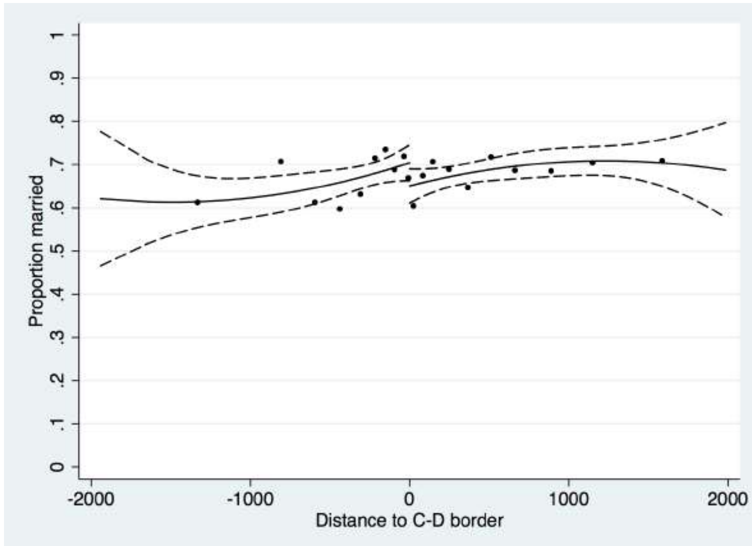
Note: Only observations within 2000 meters of the border are used. Negative distances refer to A-sided areas while positive distances refer to B-sided areas. Observations were grouped into ten bins (for each side of the border) according to distance to the border. Accordingly, each dot on the graph represents the mean proportion married for each distance-to-border decile. The solid black lines represent the predicted values of the regression of proportion married on a second-order polynomial of distance to border, where each border side has unique parameters. The dashed lines represent the 95 % confidence intervals of the fitted polynomial function.

Figure 5: Pre-treatment marriage cross-border comparisons: B-C



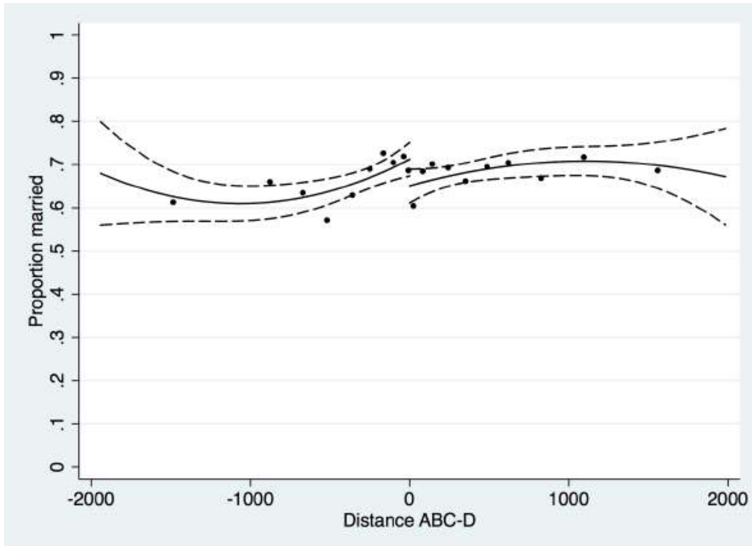
Note: Only observations within 2000 meters of the border are used. Negative distances refer to B-sided areas while positive distances refer to C-sided areas. Observations were grouped into ten bins (for each side of the border) according to distance to the border. Accordingly, each dot on the graph represents the mean proportion married for each distance-to-border decile. The solid black lines represent the predicted values of the regression of proportion married on a second-order polynomial of distance to border, where each border side has unique parameters. The dashed lines represent the 95 % confidence intervals of the fitted polynomial function.

Figure 6: Pre-treatment marriage cross-border comparisons: C-D



Note: Only observations within 2000 meters of the border are used. Negative distances refer to C-sided areas while positive distances refer to D-sided areas. Observations were grouped into ten bins (for each side of the border) according to distance to the border. Accordingly, each dot on the graph represents the mean proportion married for each distance-to-border decile. The solid black lines represent the predicted values of the regression of proportion married on a second-order polynomial of distance to border, where each border side has unique parameters. The dashed lines represent the 95 % confidence intervals of the fitted polynomial function.

Figure 7: Pre-treatment marriage cross-border comparisons: ABC-D



Note: Only observations within 2000 meters of the border are used. Negative distances refer to ABC-sided areas while positive distances refer to D-sided areas. Observations were grouped into ten bins (for each side of the border) according to distance to the border. Accordingly, each dot on the graph represents the mean proportion married for each distance-to-border decile. The solid black lines represent the predicted values of the regression of proportion married on a second-order polynomial of distance to border, where each border side has unique parameters. The dashed lines represent the 95 % confidence intervals of the fitted polynomial function.

To account for this, this study performs an analogous pre-treatment analysis that is regression-based rather than visual. In particular, it deploys a pre-treatment RDD to see if there are important discontinuities in the variable of interest. If such discontinuities are found, a more complicated approach that accounts for pre-treatment discontinuities is justified. Crucially, as implied by the previous analysis, the RDD regression allows us to account for the data set-up that the graphs may mistakenly dismiss.

In particular, the baseline estimation strategy that is used is a spatial RDD with border fixed effects, and a distance-to-border polynomial calculated at the city-border-side level. This means that we allow for spatial trends to vary within grades and cities, but also absorb the means of each specific within-city border. The estimating equation in this case amounts to:

$$Y_{ibc} = \omega_b + Grade_i + \delta_c \times Grade_i \times Dist_i + \delta c \times Grade_i \times DistSq_i + \epsilon_{ibc} \tag{2}$$

Where Y_{ibc} is the outcome for census tract i, closest to border b, in city c; ω_b is the border fixed effect; $Grade_i$ is a dummy for the grade of census tract i (this will be discussed further on in more detail); δ_c is a city fixed-effect; $Dist_i$ is the distance from the centroid of census tract i to the closest border; $DistSq_i$ is the same distance squared. Here we take a second-degree polynomial, as is standard in the literature.

Here, as for other specifications, standard errors are clustered at the city level. The justification for this clustering follows Abadie et al. (2022), where authors note that standard errors should be clustered with respect to the groups inherent to the sampling process and assignment mechanisms. In this case, the treatment is A to D assignment at the block level. With respect to the sampling process, we observe blocks that - as will be discussed later on - belonged to cities that received the maps, and more specifically, had populations of over 40,000. Thus, inclusion in our sample is determined at the city level. With respect to the assignment mechanism, blocks received "A" to "D" designation depending on the city they belonged to. For instance, a block with given average house price in New York City may received a "B" designation, but due to differing average house prices, may receive an "A" designation in Pittsburgh.

Table 2 features the RDD effects described in Figures 4 to 7, accounting for the data problems as previously outlined.

As Tables 2 shows, there are RDD discontinuities for all grade pairs. In particular, the relatively discriminated areas are worse off, as we might expact, than favoured areas. While these gaps are important evidence that justify the RDD-PSM approach, these gaps alone do not justify the method. In fact, if these gaps were to remain constant across time, an RDD-DID approach may be a better option. This assumption does not hold, if, for instance, there are other variables for which pre-treatment gaps are found. These other variables may in fact interact with exogenous shocks (e.g. wider cultural shifts in approaches to marriage), to yield "unparallel" trends between grades. Accordingly, Tables 3 and 4 feature the same pre-treatment RDD analysis, featuring other relevant

Table 2: Pre-treatment RDD cross-border comparisons (marriage)

	(1)	(2)	(3)	(4)
		Proportion married		
	A-B	B-C	C-D	B-D
RDD	0.407*	-0.0262*	0.0263*	-1.704*
	(1.13e-11)	(1.28e-12)	(3.23e-08)	(3.23e-08)
Observations	161	550	1169	354

Note: This table features simple spatial RDD effects for different border pairs. The outcome is the proportion of married individuals, prior to treatment. Here, border fixed effects and a distance-to-border second degree polynomial function, at the level of the city-grade, are used. Standard errors clustered at the city level in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

outcomes that may interact with shocks as previously described.

As Tables 3 and 4 show, there are significant differences between border sides with respect to all the featured survey variables. To minimise the effects of this, this study will not pool grades - i.e. it will only compare sequential grades, and avoid the likelihood of discontinuities, which is larger for non-sequential grade pairs (e.g. B-D).

Still, even avoiding pooling, it can not be held that the simple RDD approach adequately accounts for pre-existing gaps between grade pairs, and the key RDD assumption is simply not satisfied. To overcome this, this study deploys a spatial RDD-PSM approach proposed by Keele et al. (2015). In particular, the method entails deploying PSM and including geographical variables like border fixed effects and distances to borders.

This study deploys the Keele et al. (2015) spatial-RDD method, which is similar in spirit to the Aaronson et al. (2017) approach. In other words, it similarly augments the spatial RDD method by attributing more weight to border-pairs that are similar prior to the treatment. To this end, it deploys exact matching on border fixed effects, and standard matching on distance to border and pre-treatment covariates. Furthermore, nearest neighbour matching is used. In particular, the algorithm that is used is nearest neighbour matching.

Table 3: Pre-treatment RDD cross-border comparisons (other outcomes)

	(1) Home value	(2) Population	(3) Age	(4) Female	(5) Rent	(6) Owner-occupier
A-B	-8448.2*	-1684.2*	0.128*	-0.000833*	-7.581*	-0.0834*
	(0.000000326)	(0.000000275)	(3.19e-11)	(2.40e-12)	(5.70e-10)	(1.58e-11)
Observations	718	718	717	717	718	716
B-C	-0.0209*	687.1*	-0.107*	-0.00729*	-3.767*	-0.0209*
	(4.08e-11)	(0.000000126)	(2.58e-09)	(9.79e-12)	(4.09e-09)	(4.08e-11)
Observations	5182	5264	5227	5227	5264	5182
C-D	57.37*	-701.9*	-0.304*	0.000351*	-2.313*	-0.0332*
	(0.00000231)	(0.000000818)	(1.13e-09)	(3.61e-11)	(1.25e-08)	(3.17e-11)
Observations	9291	9291	9230	9230	9291	9170
B-D	0.711*	-48947.3*	20.18*	0.298*	-245.6*	0.711*
	(6.45e-10)	(0.0000370)	(5.90e-09)	(3.58e-10)	(0.000000230)	(6.45e-10)
Observations	323	324	324	324	324	323

Note: This table features simple spatial RDD effects for different border pairs. The outcome varies according to the column. Here, border fixed effects and a distance-to-border second degree polynomial function, at the level of the city-grade, are used. Different border pairs are considered in each row. Standard errors clustered at the city level in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

In particular, the pre-treatment variables that are used for matching mimic those found in the block-level surveys, as featured in Figure 3. These variables include: home values, population size, black population size, average age, female population, foreign-born white population, rent values, and share owner-occupiers.

Table 5 analyses pre-treatment RDD-PSM discontinuities for the outcome of interest. It is held that if this gap is null prior to treatment, then the RDD-PSM method fixes the problems outlined in Tables 3 and 4. As Table 5 shows, there are actually positive (i.e. counter to what would be expected) gaps between grades. This suggests that any post-treatment effects will be lower bounds for the true treatment effect.

2.5 HOLC grades: varying treatment intensity

The varying HOLC grades represent differing treatment intensities. This section oulines precisely what is meant by each grade. In particular, it will describe the "legend" included in the various

Table 4: Pre-treatment RDD cross-border comparisons (other outcomes) - continued

	(1) Foreign born white	(2) Black
A-B	-0.0145*	-0.0103*
	(1.14e-14)	(9.54e-15)
Observations	356	356
B-C	0.0113*	0.0556*
	(5.82e-14)	(4.79e-13)
Observations	447	447
C-dD	-0.00682*	0.0272*
	(2.42e-13)	(5.71e-13)
Observations	1036	1036
B-D	-2.205*	-0.165*
	(1.74e-09)	(2.35e-10)
Observations	324	324

Note: This table features simple spatial RDD effects for different border pairs. The outcome varies according to the column. Here, border fixed effects and a distance-to-border second degree polynomial function, at the level of the city-grade, are used. Different border pairs are considered in each row. Standard errors clustered at the city level in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$

Table 5: Pre-treatment RDD-PSM cross-border comparisons (marriage)

	(1)	(2)	(3)
	Proportion married		
	A-B	B-C	C-D
RDD-PSM	0.080***	0.134***	0.0804***
	(0.0199)	(0.0176)	(0.0199)
Observations	161	550	1167

Note: This table features RDD-PSM effects for different border pairs. The outcome is the proportion of married individuals, prior to treatment. Standard errors clustered at the city level in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

HOLC maps. Recall that these maps were covertly circulated to banks, and it was on the basis of this legend that banks were guided regarding how to approach different neighborhoods when evaluating the loan-to-value ratio (LTV).

In "A" graded neighbourhoods, banks are told to lend liberally as these are considered "hot spots". A LTV of 75-80 percent is suggested. Next, "B" areas are described as "still good" but not "hot spots", and a 65 percent LTV is proposed. On the other hand, "C" areas are described as being "infiltrated by 'lower grade' population", and lenders are told to be "conservative". Finally, "D" graded areas are describes as "hazardous" and lenders are told to refuse loans altogether.

2.6 Results

This section shows the main results for the impact of redlining on marriage in 1950, 1960, and 1970. In particular, Table 6 shows these results, with each column describing a post-treatment year, and each row describing a particular grade pair. The first row, that is the impact of B versus A grades, does not seem to show any effects for any years. This may be due to the fact that both areas were in fact favoured. The second row, C versus B, shows effects for 1950, 1960, and 1970. These effects are increasing between 1950 and 1960/1970, and show evidence of "compounding" policy effects. In particular, in 1960, C designation versus B designation decreased marriage by 0.9 percentage points. Finally, the last row, which shows effects for D versus C areas, shows effects for all years, and these too compound across time. In 1960, D designation (versus C) decreased marriage by 0.698 percentage points. In 1970, D designation decreased marriage by 1.7 percentage points.

Table 6: RDD-PSM impact of redlining areas on marriage

	(1)	(2)	(3)
	Proportion married		
	1950	1960	1970
A-B	-0.004	-0.004	-0.004
	(0.002)	(0.002)	(0.002)
Observations	3027	3354	8310
B-C	-0.006^{***}	-0.009^{***}	-0.009^{***}
	(0.001)	(0.001)	(0.001)
Observations	10535	24218	24218
C-D	-0.007^{***}	-0.017^{***}	-0.017^{***}
	(0.002)	(0.003)	(0.003)
Observations	18753	35851	35851

Note: This table features RDD-PSM effects for different border pairs. The outcome is the proportion of married individuals, post treatment. Standard errors clustered at city level in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

2.6.1 Housing credit mechanism

This section verifies whether the housing credit mechanism per se is driving the previous results. In particular, there are two mechanisms potentially yielding the results in Table 6. The first is the housing credit mechanism, which we are interested in isolating. For instance, if a couple is unable to secure housing credit, they may not get married. The second mechanism is that of neighbourhood decline. For instance, local businesses may struggle to gain access to credit, thereby leading to vacant lots and associated neighbourhood decline. Individuals in this neighbourhood may fear the increasing levels of crime, and therefore choose to not build a family or get married. This effect is proxied by house prices, as neighbourhood amenities such as crime are capitalised into house prices. In order to understand whether the results are at least partially driven by housing credit per se, Table 7 includes contemporaneous house prices in the matching process. If housing credit were irrelevant, these coefficients would drop to null significance. It is clear from the table, however, that the coefficients remain similar in magnitude and significance. It is therefore likely that the results are driven by housing credit.

Table 7: RDD-PSM impact of redlining areas on marriage (controlling for neighbourhood decline)

	(1)	(2)	(3)
	Proportion married		
	1950	1960	1970
A-B	0.011	-0.003	-0.003
	(0.007)	(0.004)	(0.004)
Observations	3027	3354	8310
B-C	-0.005	-0.011^{***}	-0.011^{***}
	(0.003)	(0.003)	(0.003)
Observations	10535	24218	24218
C-D	-0.007^{**}	-0.028^{***}	-0.028^{***}
	(0.004)	(0.003)	(0.003)
Observations	18753	35851	35851

Note: This table features RDD-PSM effects for different border pairs. The outcome is the proportion of married individuals, post treatment. All specifications include a control for contemporaneous house prices. Standard errors clustered at the city level in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

Table 8 provides an additional check for whether the credit mechanism is responsible for the area-level effects. The intuition is that credit-constrained areas are more likely to be affected by the policy, if the credit mechanism is driving results. To understand whether this is the case, the analysis assumes that D-graded observations farther from the C-D border are more credit constrained, as the tendency would be for areas to become wealthier farther from the D centroid and into higher-graded areas. Accordingly, the analysis restricts the sample to wealthier and less wealthy D-graded areas and looks at whether results are different. If the hypothesis holds, then when the sample is restricted to wealthy areas, the effects would be smaller. The first row of Table 8 features a C-D comparison where D areas are restricted to observations that are above the median distance to the C-D border (i.e. closer to the D centroid). The second row of Table 8 features the same C-D effects but with D-graded areas below the median distance to the C-D border(i.e. farther from the centroid). The first of these effects is larger in magnitude than the second, which confirms that the effects may indeed be due to the credit mechanism.

Table 8: Credit mechanism placebo test

	(1)	(2)
	Proportion married	
C-D (observations in centre of D areas)	-0.35***	
	(0.005)	
C-D (observations in outskirts of D areas)		-0.24***
		(0.024)
Observations	24797	24797

Note: This table features a heterogeneous analysis whereby in the first row the sample is restricted to D-graded observations that are 50th percentile or above distanced from the C-D border, and the second row is restricted to D-graded observations closer to the border. Standard errors clustered at the city level in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$

2.6.2 Sorting

As is shown, the policy led to a decrease in marriage, particularly in 1960 for A-B pairs, 1950 and 1970 for B-C pairs, and all years for C-D pairs. However, it is worth noting that these are area-level results, and therefore potentially the result of sorting. It is possible that the higher proportion of unmarried individuals in B-graded areas simply reflects the reshuffling of individuals across grades, rather than pointing towards the effect of housing credit on marriages. This could have happened if, for instance, unmarried individuals are more likely to rent, and B-graded areas, due to housing credit discrimination, had higher renter-occupier rates. In order to understand whether this is the case, a city-level analysis is performed. It is held that if receiving a redlining map led to a decrease in city-level marriage, then the area-level results do not merely reflect the re-shuffling of individuals.

In order to perform this analysis, county-level historical data on marriage and population is used. This data is provided by the Integrated Public Use Microdata Series' National Historical GIS (Manson et al. (2019)). Data on county-level marriage and population is used for 1950, 1960, and 1970. Furthermore, data for 1930s population is used.

In order to understand whether the area-level results are due to sorting, this study exploits quasi-randomness in treatment assignment at the city level. In particular, it exploits the fact that every city with a 1930s population above 40,000 received a map (Hillier (2005)). This clean, but arguably random, cut-off, lends itself to a standard RDD set-up. Thus, the estimating equation amounts to:

$$Y_i = \alpha + \beta_1 Cut_i + \beta_2 Pop1930_i + \beta_3 Pop1930Sq_i + \epsilon_i \tag{3}$$

Where Y_i is the number of marriages per capita in county i, Cut_i is a dummy that takes the value of one if county i had a population over 40,000 in 1930, 0 otherwise; $Pop1930_i$ and $Pop1930Sq_i$ are the 1930s population polynomial.

Table 9 features the city-level results. The first line of the table features the effect of interest - a dummy variable that takes the value of one if the city received a map. The remaining two lines feature the second-order polynomial control function of the population variable. Column 1 features results for 1950, column 2 for 1960, and column 3 for 1970. Indeed, it seems that for 1960 and 1970, the HOLC maps led to less overall marriage per capita. It is worth noting that the 1950 coefficient is close to the required significance cut-off. In sum, there there were city-level changes in marriage due to the maps. Taking the area-level and city-level results together, evidence strongly suggests that access to housing credit affects the decision to marry.

Table 9: City-level impacts of redlining on marriage

	(1)	(2)	(3)
		Proportion married	
	1950	1960	1970
1930s population cut-off	-0.00265	-0.00290^{+}	-0.00207^{+}
	(0.00170)	(0.00158)	(0.00121)
1930s population (millions)	-0.00441	-0.00762	-0.00679
	(0.00921)	(0.00852)	(0.00655)
1930s population (millions) squared	0.00140	0.00241	0.00201
	(0.00335)	(0.00310)	(0.00238)
Constant	0.0133*	0.0119*	0.0133*
	(21.62)	(21.02)	(30.74)
Observations	2997	3083	3088

Note: This table features city-level observations for 1930. It includes results for the city-level RDD where the HOLC population cut-off is the relevant discontinuity. Standard errors clustered at the city level in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, *** $p < 0.01$

2.7 Conclusion

This study analysed the extent to which obtaining housing credit affects the decision to get married. To this end, it evaluated the impact of an area-based housing credit discrimination policy. Finding that the policy's negative effects on marriage were not due to sorting or neighbourhood decline, there are some avenues for potential further work. The first stream of research that could follow from this work would be to understand the extent to which the rise in cohabitation (versus marriage) is due to rising house prices. Understanding the reasons for the rise of cohabitation are policy-relevant, as children raised in married, versus cohabiting, households, have better outcomes (Doepke et al. (2022)). Married parents invest more in their children, and are less likely to break up and consequently cause disruptions in skill acquisition. While this study found that the inability to access housing credit deters individuals from marriage, the link between trends towards cohabitation and house prices has not been established. Another, related, potential stream of work could be to understand the extent to which access to housing credit could incentivise marriage, as other benefits to marriage decrease. As education gaps between men and women narrow, so does the wage differential, and, consequently, the benefits to household specialisation.

With respect to other evidence on the impact of redlining maps, this study sought to understand whether the observed effects on marriage were due to the mortgage mechanism per se, or whether due to the documented effect on neighbourhood decline (see e.g. Krimmel (2018) or Aaronson et al. (2017)). It showed that the effects are due to the mortgages per se. Given that the mortgage mechanism is a first-order consequence of the redlining maps, and given that this mechanism affected marriage causally, it can therefore be held that the policy's final effects on outcomes like e.g. income, can at least partially be assumed to be mediated by marriage effects.

3 Covid-19 lockdowns and mental health

3.1 Introduction

Throughout the covid-19 pandemic, policymakers were faced with a bad game of "would you rather". Would you rather: contain a public health crisis? Or prevent economic collapse? Adding to this unfortunate balance were the mental health costs of pandemic containment. It became obvious that it wasn't simply a case of health versus economics, but also, that preserving physical health came with a potentially hefty mental health price tag. These effects are theoretically ambiguous - for instance, lockdowns may have spared individuals of commutes (thereby improving mental health), or on the other hand weakened social ties (thereby deteriorating mental health). Still, several studies seemed to suggest that lockdown policies, through the social isolation mechanism, eroded individuals' mental health.

However, these studies should be cautiously interpreted. To my knowledge, this literature - without exception - features limited samples from self-reported mental health, and deploy econometric methods that do not uncover causal parameters (Mautong et al. (2021), Kim & Jung (2021), Robb et al. (2020), Asiamah et al. (2021), Ma et al. (2020), Murayama et al. (2021), Killgore et al. (2020), Rauschenberg et al. (2021), Smith et al. (2020),Gonçalves et al. (2020), Lucchetti et al. (2020), Hamza et al. (2021)). This study's first research aim is to contribute to this literature by deploying larger, less selected samples, that do not rely on self-reported mental health. In particular, it deploys Twitter data in conjunction with off-the-shelf text analysis methods, to identify three particular outcomes - depression, stress, and anxiety.

A key concern in using Twitter data, however, is its representativeness on two levels. The first level is that individuals self-select into when they Tweet, so that we are only able to observe a skewed, non-random, sub-set of their moods. The second level is that Twitter users are not representative of the overall population. The first representativeness issue can be thought of as a selection problem - which, as outlined in Heckman (1979) - can in turn be thought of as a censored sample problem. While we would like to observe mood at every point in time, we only observe mood when the individual decides to Tweet. If these data points were randomly collected (e.g. the individual is randomly prompted to Tweet), the average of these moods - provided we have sufficient data points - would be close to the actual average of all moods the individual experienced throughout the day. However, the individual's decision to Tweet is endogeneous to mood. For instance, a given individual may only Tweet when experiencing extremely negative emotional states. In this case, given that less extreme and negative emotional states are not observed, our inferences concerning average mood would be downwardly skewed.

To overcome this issue, we make inferences about missing moods by leveraging insights from biology, as well as exploiting randomness in the decision to Tweet. We depart from the fact that individuals' moods react to sunlight in predictable ways (i.e. the circadian rhythm) - in other words, that in the absence of mood shocks, the functional form of mood is constant for each individual per day. Furthermore, we exploit the fact that individuals have "habitual" times that they Tweet (e.g. an individual may Tweet at breakfast every day), and that these Tweets are close to the random prompting scenario previously outlined. Given we know the functional form of moods, and given we have quasi-randomly-selected moods for each day, we can make inferences about the average mood per day.

However, even if we perfectly observed moods within our Twitter sample, problems of external validity would persist, given that Twitter users are not representative of the overall population. To overcome this, the study adjusts area-level Twitter scores to match distributions of a representative mental health index.

This Twitter data is then used to answer questions about the mental health impacts of the pandemic. To this end, the UK's pandemic containment policies are evaluated. The UK government pursued a first national lockdown, whereby all regions were equally restricted, Then, it pursued a tiered lockdown, whereby regions were attributed different levels of restriction. Finally, it pursed a second national lockdown, where regions were once more equally restricted. This study is interested in understanding whether the lockdown policies affected individuals' moods. This larger question can be broken down into three main sub-questions. The first is whether the second national lockdown - compared to a similar period in pre-pandemic year altered mental health. The second question is whether severity of social isolation affected mental health. The third question is whether the lockdowns had mental health potentiation effects - i.e. whether experiencing more punitive social isolation in a previous lockdown period led to a larger mental health shock in a subsequent lockdown period.

The first question - of whether the second national lockdown affected mental health vis-a-vis the same period in a pre-pandemic year - entails comparing the lockdown period (here we perform this analysis for the second national lockdown) with the same period in a pre-pandemic year. Here, time trends are accounted for.

The second question - of whether severity of social isolation affected mental health - is tackled by exploiting the UK's "tier" policy. In sum, different regions were assigned different severity of social isolation, on a scale of tier one (least severe) to tier three (most severe). Importantly, these tiers were updated almost daily. This study uses user-level fixed effects and essentially compares, e.g., Tier one areas to tier two areas, given that Tier one areas eventually became tier two areas.

The final question - of whether there were mental health potentiation effects between lockdowns - is tackled by using the method outlined for question one, but for the second national lockdown. To understand whether there were potentiation effects, an interaction effect for the average tier of the region in the previous regional lockdown is deployed.

In particular, this study finds that the second national lockdown increased stress and anxiety, but had no effect on depression. Furthermore, it finds that more punitive social isolation has no effects on depression or stress, but increases anxiety. Finally, this study finds no potentiation effects of mental health states throughout the pandemic.

3.2 Data

This study uses two main sources of data. The first pertains to the outcome variable - mental health. To this end, geo-located Twitter data is used. The second pertains to the independent variable - tier status. To this end, UK legislation was manually scraped. Ensuing paragraphs describe each of these in turn.

3.2.1 Tweet collection

In order to understand the effect of the national lockdowns on mental health, it is necessary to have Tweets for the UK. As will be further explained, in order to account for Tweets' lack of mood representativeness, it is also necessary to have several Tweets per user. Finally, in order to understand the effects of the tiered lockdowns on mental health, it is necessary to have Tweets that are geo-located at the Local Authority District (LAD) level. In sum, the analysis requires that we have several Tweets per user, and that we know where LAD each user belongs to.

In a first step, all Tweets written in the UK during the second national lockdown period were collected.[1] In a second step, UK-based users were identified from the sample collected in the first step. In a third step, the timelines of these users were collected for relevant periods. In particular, I download Tweets for the tiered lockdown period, the second lockdown period (November 5th-December 2nd), and all relevant analogous period Tweets for 2017 and 2018. In a fourth step, it was necessary to, within the UK, establish which LAD the user belonged to. Here, it is worth noting that there are two types of geo-located Tweet. In the first, users opt-in to disclosing the precise coordinates of all their Tweets. In the second, users "tag" the location that they are in for a given Tweet. It is estimated that 0.71 percent of users enable precise geo-location, and 2.25 percent of Tweets use broader geo-location (Huang & Carley (2019)). In order to convert broadly geo-located tweets (e.g. "Liverpool") into LADs, I first obtained precise coordinates for the broad location. To this end, I used the Google Maps API, which converts addresses into the coordinates of the centroid of the location in question. With the coordinates of each Tweet, I then assign each coordinate an LAD by overlaying the points with a polygon shape file of LADs.

Many users only enabled geo-location temporarily, so that the LAD identified in the first wave of data collection (i.e. downloading all Tweets in the UK for the second national lockdown period)

[1] In order to download these Tweets, it was necessary to connect to the Twitter Application Programming Interface (API) This API essentially serves as a bridge between Twitter data and the individual's personal computer. Prior to August 2020, there were only three types of APIs that individuals could access: standard (free), premium (self-serve paid), and enterprise (custom paid). Individuals with standard access were unable to download historical or archived Tweets, unless they had specific usernames in mind. From August 2020, however, Twitter developed the Academic Research API that allowed researches to access the full archive of Tweets. Accordingly, this study makes use of this API. In order to download the Tweets, it uses the twarc2 command line tool and Python library. It is also worth noting that this search query does not necessarily download all the Tweets that were written in the UK during the time frames, because only a sub-sample of users choose to geo-locate their Tweets.

is the only location that is available. It is therefore assumed that the geo-located Tweet in the first stage of data collection is representative of where the individual was located throughout the tiered lockdown, and that the individual remained in the UK throughout the second national lockdown.

3.2.2 Text analysis

With a series of relevantly geo-tagged tweets for the necessary time frame, it is necessary to "mine" the text/content of the Tweets for mood. This study follows off-the-shelf methods for achieving this, and provides a brief literature review of the state of the art on mood inference from Tweets. It will then outline the strategy herein used, as well as show how successful the method was.

Simple text analysis methods can be thought of as the same as standard quantitative methods, but where each variable X is a word. Although there are more complicated methods, a simple "bag-of-words" approach suffices for the purposes of this study. Machine learning methods essentially uncover the parameters inherent to an X-Y relationship, then use these parameters to predict Y, given X.

Crucially, then, it is necessary to have an X-Y set. In this case, this would be a set of Tweet words associated with a given emotional state. Here, the literature is divided into three main streams with respect to how this is achieved. The first set of studies use human annotation - that is, humans read Tweets and label them with the emotion they think is latent. This approach has obvious pitfalls - it is cumbersome to deploy, but more crucially, it is not obvious that a human can successfully gauge another's emotion by looking at a Tweet. This is further exacerbated by the fact that social media is a panopticon where individuals mold their behavior to craft facades and identities. In short, it is possible that an individual appears to be one way but feels another. However, this method is popular, particularly in earlier studies: Roberts et al. (2012), Bravo-Marquez et al. (2013), Bosco et al. (2013), Resch et al. (2015), Mohammad & Bravo-Marquez (2017), Mohammad & Yang (2011), Pestian et al. (2012), Strapparava & Mihalcea (2008), Coban et al. (2018), Kalamatianos et al. (2018), Aloufi & El Saddik (2018), Agarwal et al. (2011), An et al. (2014), Asiaee T et al. (2012), Aston et al. (2014), Bakliwal et al. (2012), Barbosa & Feng (2010), Bora (2012), Curini et al. (2015), Burnap et al. (2015), Homan et al. (2014), Huang et al. (2014).

Other studies rely on the self-disclosure of certain emotions. The first set of studies relies on self-disclosure of underlying illnesses in e.g. Twitter biographies (see Birnbaum et al. (2017), Coppersmith, Dredze, Harman & Hollingshead (2015), Coppersmith et al. (2014), Coppersmith et al. (2016), Loveys et al. (2017), McManus et al. (2015), Mitchell et al. (2015), Resnik et al. (2015), Saha et al. (2017), Coppersmith, Dredze, Harman, Hollingshead & Mitchell (2015)). While this method might be useful for understanding the extent to which individuals might suffer from a mental illness, it is less useful for understanding the latent emotions in their Tweets.

Finally, the last category of studies uses self-disclosed emotions at the Tweet (rather than biography/person) level. How they do this takes two forms. The first form entails inferring emotions from emojis used in Tweets (see Pak & Paroubek (2010), Aono & Himeno (2018), Aisopos et al. (2011)). The second form entails exploiting hashtags (e.g. "#feelinglow"). This study follows the hashtag studies, and is therefore aligned with Barbieri & Saggion (2014), Janssens et al. (2014), Mohammad (2012), Sintsova et al. (2014).

Thus, the training set was comprised of Tweets with associated hashtags. It is worth discussing the specific form of the dependent variable, as it is not obvious how emotions should be categorised. While there are several theories on how to label and categorise emotions, in a similar vein to Kim et al. (2012), Suttles & Ide (2013), Mohammad & Yang (2011), this studies draws from Plutchik's eight bipolar emotions. In particular, this theory holds that every emotion has an equal and opposite emotion. In contrast to Plutchik, this study does not focus on emotions per se, but rather, "collections" of emotions. In this study, I focus on three emotional states of interest - depression, anxiety, and stress. It is held, for instance, that "stress" is an umbrella term for several negative/positive emotions (Folkman (1984)).

In particular, bearing in mind Plutchik's concept of bipolar emotions, the dependent variable is a binary variable that takes the value of one if the individual is experiencing one of the aforementioned emotional states, and zero if the individual is experiencing the opposite emotional state. In particular the following emotional state pairs were identified: anxious-relaxed; stressed-calm; depressed-happy.

To obtain this training dataset of three dichotomous emotions, this study downloaded sets of hash-tagged Tweets. For instance, the "depression" dataset was half comprised of Tweets with "#depressed" and half comprised of tweets with "#happy".

The parameters are uncovered using a subset of the this data. The remainder of the data is used to test the extent to which the model was successful. For each of these six categories, 10,000 Tweets were downloaded, with the training set consisting of a sub-sample of 7,500 Tweets, and the test set consisting of the remaining 2,500 Tweets. Table 10 presents how successful this training was. That is, it trains the Naive Bayes algorithm on a sub-set of the training sample, then predicts outcomes for the remaining training set. Given that we know actual outcomes for this remaining sub-set, the predictions herein obtained give us a sense of how accurate the method is. The table features. the method's precision,[2] recall,[3] and F1.[4] Clearly, the approach is rather successful, with F1 scores ranging from 95.28 to 99.64.

It is worth noting that the Naive Bayes method is only able to predict binary outcomes, which means that Tweets will be assigned an emotional extreme within the dichotomous emotion spectrum.

Table 10: Confusion matrix

	Depression	Anxiety	Stress
Precision	96.43	94.50	97.50
Recall	99.43	96.07	98.72
F1	99.64	95.28	98.11

Note: Numbers expressed as a percentage.

3.2.3 Censored sample within users

As noted in the introduction, we only observe mood when the individual decides to Tweet. In effect, this can be thought of as a selection problem - which, as outlined in Heckman (1979) - can in turn be thought of as a censored sample problem. Censored samples and strategies with how to deal with them were first introduced in Tobin (1958), where it is argued that it is problematic when we have a variable that takes a single value if its true value goes beyond/below a certain

[2] True positives over true positives plus false positives.

[3] True positives over true positives and false negatives.

[4] A "harmonic mean" of precision and recall, which is the product of both, divided by the sum of both, multiplied by two.

threshold. In this case, the goal is to infer average daily mood from Tweets. In an ideal scientific set-up, we would observe momentary moods, cluster moments meaningfully, and take averages to infer the overall average mood. Approximations to this include the daily reconstruction method where individuals are asked to reflect on previous moods (Kahneman & Krueger (2006)) - given this method's reliance on skills that are endogeneous to mood itself, this is a problematic approach. An approximation to the ideal setting is mappiness data (MacKerron & Mourato (2020)), which randomly prompts individuals to state their mood. In addition to standard issues pertaining to self-report bias, this method of data collection is not scaleable.

Recall, ideally we would want to track every mood at every point of a meaningful time frame. With Twitter data, we have only a sub-sample of these points. To reiterate, individuals choose when to Tweet, and are not randomly prompted to Tweet. If this decision was orthogonal to the outcome (mood), it would not pose a problem. However, decisions are often emotionally-based, and the decision to express oneself in a micro-blog is doubtlessly emotionally charged.

This is problematic for two reasons. The first is that our data is censored in that we only observe a few elements of the distribution, which would render our parameters/moments uninformative if our sample is small (at the individual-day level). The second issue is that in addition to only observing a sub-sample of Tweets, the sub-sample we observe is selected in a way that is problematic. This selection problem renders β_1 biased due to the fact that the policy variable - i.e. the allegedly exogenous shock - possibly determines the decision to Tweet. In other words, the measurement error in Y_{it} is correlated with the policy variable. An example of this might be that as lockdown progresses, depressed individuals lose the motivation to Tweet. In this example, the parameter would be upwardly biased, because only happier people are left in the sample.

This study overcomes this issue by imputing missing Tweets. To do so, it departs from the idea that the solar cycle influences mood via the circadian rhythm - that is, our "inner clock" that adapts physiology to different times of the day (Watson & Clark (1997)). Crucially for this study, each individual has a certain in-built reaction to daylight, and this reaction regulates mood. In fact, a survey of Twitter users shows that this theory holds (Dzogang et al. (2017)). This circadian-rhythm-defined mood trajectory can be thought of as the mood trajectory of the individual if he were not exposed to any wellbeing shocks. This is impossible, as the world is constantly in flux. Rather, it can be thought of as a Zen Buddhist's reaction to wellbeing shocks - i.e. precisely no reaction whatsoever.

The basic intuition for the method of imputing Tweets is that the functional form of this mood-pattern, for each individual, is uncovered through the data by obtaining the average moods for each daily hour across all days that we observe data. It is held that for each hour of the day, based on sunlight, the individual has a "baseline" mood. The intuition for averaging all days is that the shocks that the individual experiences, which vary by day and hour, have mean zero. Thus, by obtaining the average of the hourly mood across all days, we are left with the baseline circadian rhythm.

In principle, this functional form can be derived by regressing mood on a flexible polynomial of day-hour, for each individual. If the residual is random, and zero on average, we'd arrive at the circadian rhythm-defined fluctuation of mood. However, as noted, individuals choose when to Tweet, rending the residual non-random and likely not mean zero. For this reason, we will derive our functional form using only a sub-set of Tweets that are not the product of prior moods. We refer to these Tweets as "habitual", due to the fact that they are written, as label would suggest habitually, rather than due to mood-derived impulses. An example of a habitual Tweet would be if one checked Twitter, religiously, every day during breakfast. In order to identify these habitual Tweets, we will understand if there are significant patterns in the hours at which a person Tweets.

Once we have identified the functional form of the circadian-rhythm-determined mood for each user, we assume that this mood trajectory is constant. Accordingly, changes in average mood per day can be explained by differing function intercepts, which are in turn obtained by taking the average of actual, but habitual, tweets, at the day level. To reiterate, in effect, this constant trajectory gives us the individual's mood response to sunlight. While this is constant, the individual may wake up with different "intercepts" (i.e. baseline moods) each day.

In order to implement this, we can take the following strategy. We aggregate raw mental health at the day x hour x individual level and regress this on a flexible polynomial of hour, as well as the average Tweets produced per day, and individual x day fixed effects. The fixed effects give us the baseline mood at the beginning of the day for each individual. These fixed effects are then used as the dependent variable in our main regressions.

In sum this amounts to:

$$Y_{ith} = \theta_{it} + \beta_1 TweetsPerDay_i + \beta f(hour_{ih}) + \epsilon_{ith} \tag{4}$$

Where Y_{ith} is the mental health of individual i on day t and hour h; $TweetsPerDay_i$ is the average volume of Tweets per day for individual i; f($hour_{ih}$) is a polynomial of a dummy that takes the value of one if the hour is equal to h for individual i. We then use θ_{it} as the outcome in our main regressions, and it represents the mental health of individual i at day t, adjusted for the fact that Tweets are not representative of average daily moods.

3.2.4 Representativeness of sample

Another key challenge to using Twitter data is the representativeness of the sample. It is possible, and likely, that the average Twitter user does not represent the UK population overall. To overcome this, this study weights the sample in such a way that mimics the mental health of the UK as shown by a representative dataset.

In particular, it exploits the fact that the Small Area Mental Health index (SAMHI) (Daras & Barr (2020)) is representative at the lower super output (LSOA) level. Given that LSOAs are more or less constant in population, this dataset provides an overall view of how mental health, and in particular depression, is ranked across the UK. The SAMHI combines data on mental health from multiple sources (NHS-Mental health related hospital attendances, Prescribing data – Antidepressants, QOF - depression, and DWP - Incapacity benefit and Employment support allowance for mental illness) into a single index. They take several different measures of mental health - for instance mental health related hospital admissions or quantity of antidepressants prescribed - and divide them by the total LSOA population. They then produce an index using these different measures. The SAMHI data is calculated at LSOA x year level - in other words, the data provides one data point per LSOA per year. This frequency is insufficient for our analysis, which requires daily tracking of mental health throughout the lockdown periods.

The overall strategy that is used to weight the sample is as follows. First, day-level Twitter depression for each user (as previously described in imputation method) is aggregated at the LSOA level. Worth noting that this is done for pre-pandemic years, given that the mental health measures that SAMHI uses likely suffer from lags (e.g. mental health hospitalisations during the pandemic most likely don't reflect the mental health toll of the pandemic). In a second step, LSOAs are ranked by Twitter depression. In a third step, LSOAs are similarly ranked per SAMHI. Each Twitter LSOA is then attributed the weight of inverse distance to SAMHI ranking. That is, LSOAs for which the Twitter ranking perfectly matches the SAMHI ranking are given a weight of one.

For instance, suppose that there are five districts. District A ranks first for depression according to SAMHI, but ranks fourth for depression according to the Twitter data. Then suppose District B ranks second for depression according to SAMHI, and also ranks second for depression according to the Twitter data. Given that district B's Twitter score is closer to the score given by the representative SAMHI scale, the observations located inside it would receive a higher weight than those located in district A.

3.3 Policy

This section describes the UK government's Covid-19 lockdown policy in further detail. These policies were formalised as statutory acts, with regulation aiming at curbing the coronavirus' spread. These policies were implemented from March 21st 2020 onwards. Prior to this date, members of the public were merely given guidance. The first of "The Health Protection (Coronavirus, Restrictions) (England) Regulations 2020" enforced closures of businesses that posed a high risk of contagion. These regulations lasted five days, after which, on March 26th, a nation-wide lockdown was enforced. Amongst many measures, individuals were barred from leaving their homes unless they had a "reasonable excuse". These restrictions were progressively eased into the summer, and subsequently, nation-wide policies were replaced by ad-hoc local policies.

As infection rates worsened, on October 14th, these local policies were replaced by a formal tier system, which were deployed at the district council level. In particular, there were three tiers, with tier one facing the most relaxed restrictions, and tier three facing the most severe. This section will outline crucial differences between tiers, rather than exhaustively transcribe every element of the statute.

Tiers differed with respect to the extent to which gatherings were allowed. Indoor gatherings of up to six people were allowed in tier one zones, but not in tier two or three. Individuals in tier one and two were allowed to gather in groups of up to six in private outdoor areas, while group three individuals were not. Still, all tiers allowed for gatherings of up to six people in public outdoor areas.

With respect to businesses that were open, in all areas, restaurants and bars closed at 10 pm. In tier three areas, alcohol was only allowed to be served if accompanied with food.

This tiered lockdown period, known as phase one, lasted from October 14th to November 5th. On November 5th, another national lockdown was instated until December 2nd 2020. On December 5th, the tier system was reintroduced. On January 6th, a final national lockdown was introduced. The period of analysis for this study is the first tiered lockdown period (October 14th to November 5th) and the second national lockdown (November 5th-December 2nd 2020).

In first instance, an analysis of the regional tiered lockdown will be performed. This analysis will allow us to understand how lockdown severity deferentially affected mental health. As will be discussed, crucially for the estimation strategy, the allocated tiers were regularly updated. In practice, restrictions tended to increase in severity - i.e. areas only moved down tiers. Table 11 presents the timings for these switches. Column two features the date of the change, column two features the number of the statutory act, and finally, the ensuing three columns feature which county districts were placed under which tiers.

Table 11: Tier lockdowns

Date	Legislation	Tier 3	Tier 2	Tier 1
October 14th	1103		Ashfield, Barnsley, Bassetlaw, Birmingham, Blackburn with Darwen, Blackpool, Bolton, Bradford, Broxtowe, Burnley, Bury, Calderdale, Cheshire East, Cheshire West and Chester, Chorley, County Durham, Darlington, Doncaster, Fylde, Gateshead, Gedling, Hartlepool, High Peak, Hyndburn, Kirklees, Lancaster, Leeds, Leicester, Manchester, Mansfield, Middlesbrough, Newark and Sherwood, Newcastle upon Tyne, North Tyneside, Northumberland, Nottingham, Oadby and Wigston, Oldham, Pendle, Preston, Redcar and Cleveland, Ribble Valley, Rochdale, Rossendale, Rotherham, Rushcliffe, Salford, Sandwell, Sheffield, Solihull, South Ribble, South Tyneside, Stockport, Stockton-on-Tees, Sunderland, Tameside, Trafford, Wakefield, Walsall, Warrington, West Lancashire, Wigan, Wolverhampton, Wyre	Rest of England
October 17th	1131	Blackburn with Darwen, Blackpool, Burnley, Chorley, Fylde, Hyndburn, Lancaster, Pendle, Presto, Ribble Valley, Rossendale, South Ribble, West, Lancashir, Wyre		
October 17th	1128		Barking and Dagenham, Barnet, Basildon, Bexley, Braintree, Brent, Brentwood, Bromley, Camden, Castle Point, Chelmsford, City of London, Colchester, Croydon, Ealing, Elmbridge, Enfield, Epping Forest, Greenwich, Hackney, Hammersmith and Fulham, Haringey, Harlow, Harrow, Havering, Hillingdon, Hounslow, Islington, Kensington and Chelsea, Kingston upon, Thames, Lambeth, Lewisham, Maldon, Merton, Newham, Redbridge, Richmond upon Thames, Rochford, Southwark, Sutton, Tendring, Tower Hamlets, Uttlesford, Waltham Forest, Wandsworth, Westminster, York	
Otober 22nd	1154	Barnsley, Bolton, Bury, Doncaster, Manchester, Oldham, Rochdale, Rotherham, Salford, Sheffield, Stockport, Tameside, Trafford, Wigan	Coventry, Slough, Stoke-on-Trent	
October 27th	1176	Ashfield, Bassetlaw, Broxtowe, Gedling, Mansfield, Newark and Sherwood, Rushcliffe		
October 30th	1183		Amber Valley, BolsoverCannock Chase, Charnwood, Derby, Derbyshire Dales, Dudley, East Riding of Yorkshire, East Staffordshire, Kingston upon Hull, City of Lichfield, LutonNewcastle-under-Lyme, North East Lincolnshire, North Lincolnshire, Oxford, South Staffordshire, Stafford, Staffordshire Moorlands, Tamworth, Telford and Wrekin	
November 2nd	1192		Shropshire	

Finally, the second national lockdown's effects on mental health will be analysed. In addition to understanding the second lockdown's effects, a heterogeneous analysis that compares the effects of the second lockdown on residents that experienced different lockdown intensities in the regional tiered lockdowns will be performed. Essentially, this analysis allows us to understand whether mental health shocks interact with prior shocks. In other words, whether, e.g. hardship experienced in the past renders individuals' mental health more vulnerable to future shocks.

3.3.1 Lockdown tiers

This study considers all lockdown periods in the UK until December 2nd 2020. In order to obtain dates for these lockdowns, archives from the House of Commons Library were accessed. While obtaining these dates was fairly straightforward, obtaining data for the regional tiered lockdowns (October 14th to November 5th 2020) required manually scraping UK legislation. In particular, the UK legislation archive was used, and the terms "The Health Protection (Coronavirus Restrictions)" deployed in the search. Each statute/amendment outlines which areas changed tiers, and to which tier they moved, vis-a-vis the latest statute/amendment.

3.4 Theoretical framework

This study is an impact evaluation of lockdown policies during the covid-19 pandemic. In particular, part of the analysis will seek to understand the functional form of the mental health states cause by successive pandemic shocks. In other words, following the psychology literature, whether the effects of the pandemic on mental health were additive or multiplicative (Lepore & Evans (1996)). To illustrate these concepts, suppose an individual experiences two stressful events within the same week. The additive stress hypothesis holds that the total stress caused by these two events would be the same regardless of them occurring in the same week - in other words, that stressors and their impact on stress are orthogonal to each other. On the other hand, the multiplicative hypothesis holds that these two reactions are not orthogonal, and that the second stressful event may either catalyse more stress (potentiation effects) or less stress (attenuation effects) than if the first event had not preceded it.

Psychology literature from the 1990s generally agreed that effects of stressors were not orthogonal, but whether subsequent stressors potentiate or attenuate prior stressors is still contested. For instance, Myrtek & Spital (1986) find that in a lab setting, the sum of reactions to shocks did not exceed that of reactions to the most acute stressors. This is suggestive of multiplicative attenuation effects. Similarly, McGonagle & Kessler (1990) find that individuals tend to get accustomed to stress, so that future stressors yield fewer effects. In particular, they found that chronically stressed individuals were less responsive to acute life stressors than not chronically stressed individuals. In contrast, some studies find that there are potentiation effects. For instance, Caspi et al. (1987) find that air pollution caused more stress in individuals who experienced more stressful life events.

Part of this lack of consensus may have been due to the fact that this literature had several limitations, particularly with respect to validity. Studies performed in lab settings may have been internally valid, but ultimately lack external validity due to sample but mainly due to their lack of real-world relevance. On the other hand, studies that feature "in the field" research typically suffered from internal validity as individuals self-select into, e.g., stressful life events. More recent literature has dissected the mechanisms that may lead to either attenuation or potentiation of mental health states. For instance, Hartig et al. (2003) find that blood pressure decreases more rapidly after a stressful event if the surroundings the individual is located in are aesthetically pleasing. This study points to the fact that mental health shocks are not orthogonal to each other, and that there are potentiation effects.

Whether or not these potentiation effects held for the coronavirus pandemic is a concern of this study. As will be further explained in the estimation section, part of this analysis entails understanding the effects of the second lockdown, according to the severity of lockdown experienced previously. This leverages a unique feature of UK coronavirus policy, whereby certain regions received more punitive measures, followed by blanket national lockdowns.

3.5 Estimation Strategy

This section outlines the estimation strategies that allow us to uncover the effects of different stages of the pandemic.

The first goal of the estimation is to understand if being in a more stringently locked-down tier had negative effects on mental health. Thus, the simple OLS specification amounts to:

$$\hat{\theta_{ilt}} = \alpha + \beta_1 T2_{lt} + \beta_2 T3_{lt} + \epsilon_{ilt} \quad (5)$$

Where $\hat{\theta_{ilt}}$ is the adjusted average mental health of individual i in local authority l at day t as outlined in equation (1); $T2_{lt}$ is a dummy that takes the value of 1 if local authority l was tier two at day t; $T3_{lt}$ is a dummy that takes the value of 1 if local authority l was tier three at day t.

However, β_1 and β_2 may be biased due if there is a non-random allocation of tiers, and the variables that determine tier allocation also affect mental health. In fact, tier allocation was based on data regarding, for instance, case numbers and casualties (the precise formula, however, is not known). These coronavirus-related variables may in turn be related to other area-level variables that are related to individual-level determinants of mental health. For example, case numbers may have been higher in areas where individuals tend to use public transportation. These cities may be larger, and may be inherently more stressful places to live. In this case, it is the city size that is affecting mental health, rather than the pandemic per se.

Accordingly, the main specification controls for fixed effects at the level of the individual. By including these fixed effects, variation stems from individuals who experienced several different tiers.

Thus the main estimation strategy amounts to:

$$\hat{\theta_{ilt}} = \theta_i + \beta_1 T2_{lt} + \beta_2 T3_{lt} + \epsilon_{ilt} \quad (6)$$

Where $\hat{\theta_{ilt}}$ is the adjusted average mental health of individual i in local authority l at day t as outlined in equation (1); θ_i is a individual-level fixed effect; $T2_{lt}$ is a dummy that takes the value of 1 if local authority l was tier two at day t; $T3_{lt}$ is a dummy that takes the value of 1 if local authority l was tier three at day t.

Finally, we are interested in understanding the second lockdown's effects on mental health. In sum, the second lockdown period is compared to the pre-pandemic period, which is assumed to be 2018 rather than 2019, as the pandemic started at the end of 2019. Thus the equation amounts to:

$$\hat{\theta_{ilt}} = \alpha + \beta_1 LD2_t + \beta_2 LD2Day_t + \beta X_t + \epsilon_{ilt} \tag{7}$$

Where $\hat{\theta_{ilt}}$ is the adjusted average mental health of individual i in local authority l at day t as outlined in equation (1); $LD2_t$ is a dummy that takes the value of one if day t belongs to the second lockdown period (November 5th-December 2nd 2020) and zero if it belongs to the same period in the pre-pandemic year of 2018; $LD2Day_t$ is a variable that counts the duration of the lockdown period for the pandemic year (e.g. $LD2Day_t$ will equal one on the 2nd of December 2020, two on the 3rd of December 2020, and so on) and takes the value of zero for pre-pandemic years; X_t is a vector of day dummies.

The second national lockdown allows us to understand whether mental health shocks are correlated with one another. More specifically, it is possible that successive shocks lead to progressively larger mental health impacts. In order to understand whether this is the case, this study exploits the large exogenous shock on mental health that is the second UK national lockdown. It therefore interacts this second lockdown (as before - both in dummy form and in "days of lockdown" form) with the average previous tier in the regional lockdown. This latter variable is calculated in the following way:

$$AverageTier_l = \frac{\sum_{t=1}^{N} T_l}{N} \tag{8}$$

Where $AverageTier_l$ is the average tier level of local authority l; N is the total duration of the tiered lockdown; T_l is the tier level for local authority l at time t.

With this variable, we are able to calculate heterogeneous effects for the second lockdown, in similar spirit to equations (1) and (2). Thus the equation amounts to:

$$\hat{\theta_{ilt}} = \alpha + \beta_1 LD2_t \times AverageTier_l + \beta_2 LD2Day_t + \beta_3 AverageTier_l + \beta_4 X_t + \epsilon_{itl} \tag{9}$$

Where $\hat{\theta_{ilt}}$ is the adjusted average mental health of individual i in local authority l at day t as outlined in equation (1); $LD2_t$ is a dummy that takes the value of one if day t belongs to the second lockdown period (November 5th-December 2nd 2020) and zero if it belongs to the same period in the pre-pandemic year of 2018; $AverageTier_l$ is the average tier of local authority l; $LD2Day_t$ is a variable that counts the duration of the lockdown period for both pandemic (e.g. $LD2Day_t$ will equal one on the 2nd of December 2020, two on the 3rd of December 2020, and so on) and takes the value of zero for pre-pandemic years; X_t) is a vector of day dummies.

3.6 Results

Tables 12 and 13 present the descriptive statistics. These are the raw mental health scores, that is, without accounting for selection into Tweeting or LAD weight. Table 12 features mental health outcomes for the second national lockdowns. The first column refers to the the pre-pandemic period - that is, the same months/days as the lockdown period, but for the year prior to the pandemic (2018). Relative to the analogous period before the pandemic, stress and anxiety seem to have increased during the second lockdown period, while depression remained roughly the same.

Table 13 describes mental health outcomes for the different Tiers, during the tiered lockdown period. Depression is highest for tier two LADs, while tiers one and two have the highest stress levels. Finally, the first tier has the highest anxiety level. These counter-intuitive differences between tiers may be explained by the endogeneity of tier status, which will later be accounted for.

Table 12: Descriptive statistics: second national lockdown

	Second lockdown pre	Second lockdown post
Depression	0.74	0.74
	(0.43)	(0.43)
Stress	0.65	0.66
	(0.47)	(0.47)
Anxiety	0.47	0.50
	(0.49)	(0.49)
Users	12,153	8,043
Users*Tweets	78,849	134,490

Note: The first three sets of two rows feature the mean and standard deviation (in parentheses) for each outcome.

Table 13: Descriptive statistics: tiered lockdown

	Tier 1	Tier 2	Tier 3
Depression	0.74	0.75	0.75
	(0.43)	(0.43)	(0.42)
Stress	0.50	0.51	(0.52)
	(0.49)	(0.49)	(0.49)
Anxiety	0.66	0.67	(0.68)
	(0.47)	(0.46)	(0.46)
Users	5,766	7,779	4,529
Users*Days	35,217	47,895	31,094

Note: The first three sets of two rows feature the mean and standard deviation (in parentheses) for each outcome. Columns represent tiers. Each LSOA can belong to several different tiers.

Table 14 features the effects of the regional tiered lockdown, referring to equations (2) and (3). The main goal of this table is to understand whether slight differences in levels of social isolation contribute to mental health outcomes. Columns 1-2 refer to the depression outcome, columns 3-4 refer to the stress outcome, and columns 5-6 refer to the anxiety outcome. Within each outcome, each column presents a different specification. Even columns feature the simple OLS impacts of being tier two or tier three vis-a-vis tier one. Odd columns feature the individual-level fixed effects specification. This section will describe the results for each mental health outcome in turn. With respect to the tiered lockdown's effects on depression, the OLS specification suggests that, compared to tier one, tier two led to a 2.7 percentage point decrease in depression. However, this effect becomes null in the more robust user-fixed effects specification. This difference between the biased (OLS) and more robust (fixed effects) parameter suggests that individuals that resided in areas with tier two designation - compared to tier one areas - were on average less depressed, due to reasons other than the lockdown per se. With respect to stress, OLS effects suggest that tier three areas were more stressed than tier one areas by 1.48 percentage points. Given that the more robust fixed effects parameter is null, this allows us to infer that tier three areas are on average more stressed than tier one areas, due to reasons unrelated to the lockdown per se. Finally, with respect to anxiety, OLS effects show no difference between tier two and tier one areas. However, in the more robust fixed effects specification, tier two designation increased anxiety by 1.57 percentage points. The lack of OLS effect and positive fixed effects effect suggests that tier two areas, compared to tier one, were likely less anxious due to reasons other than the lockdown per se. OLS effects suggest that tier three designation, compared to tier one, increased anxiety by 1.48 percentage points. These effects become larger in magnitude in the more robust fixed effects specification, where tier three designation increases anxiety by 3.7 percentage points. Here, differences between OLS and fixed effects parameters suggest that individuals in tier three areas were less anxious than tier one areas due to reasons other than the lockdown per se.

Table 14: Impact of tiered lockdown on mental health

	(1)	(2)
	Depressed	
Tier 2	-0.0270**	0.00293
	(-2.38)	(0.31)
Tier 3	-0.0127	0.0113
	(-1.19)	(0.91)
User fixed effects	NO	YES
Observations	94974	94974
	(3)	(4)
	Stressed	
Tier 2	0.00399	0.00159
	(0.40)	(0.11)
Tier 3	0.0148**	0.00179
	(2.02)	(0.12)
User fixed effects	NO	YES
Observations	94974	94974
	(5)	(6)
	Anxious	
Tier 2	-0.00454	0.0157***
	(-0.53)	(3.39)
Tier 3	0.0164*	0.0370***
	(1.94)	(2.69)
User fixed effects	NO	YES
Observations	94974	94974

Note: This table refers to the impact of Tier 2 and Tier 3 (dummy variables) designation on mental health, with Tier 1 as the base category. Columns 1-2 refer to effects for depression, columns 3-4 refer to effects for stress, and columns 5-6 refer to effects for anxiety. Odd columns feature baseline OLS effects, while even columns feature user-level fixed effects. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tables 15 and 16 present the impact of the second lockdown on mental health outcomes, referring to equation (4). The main goal of this table is to understand whether a mental health shock interacts with previous shocks. To this end, we consider the interaction of the second lockdown with the severity of the tiered lockdown, in line with equation (6). It is held that if this interaction term is significant, then there were mental health potentiation effects during the pandemic lockdowns. Odd columns tell us the impact of the second lockdown vis the same period in the previous year. This paragraph will consider each of the outcomes in turn. With respect to depression, there were no changes in the second lockdown period, vis-a-vis the analogous period prior to the pandemic. In contrast, the second lockdown period increased stress by 4.67 percentage points. However, there were no potentiation effects. Finally, the second lockdown period increased anxiety by 3.67 percentage points. Similarly to other outcomes, there were no potentiation effects for the anxiety outcome.

Table 15: Impact of second national lockdown on mental health

	(1)	(2)
	Depressed	
Days second lockdown	0.000120	0.000133
	(0.000405)	(0.000405)
Second lockdown	0.00836	0.0153
	(0.00535)	(0.0102)
Average tier		-0.00744
		(0.00501)
Second lockdown × Average tier		-0.00371
		(0.00509)
Constant	-0.0225**	-0.00979
	(0.0101)	(0.0133)
Day fixed effects	YES	YES
Observations	177074	177074
	(3)	(4)
	Stressed	
Days second lockdown	-0.000342	-0.000352
	(0.000340)	(0.000341)
Second lockdown	0.0368***	0.0467**
	(0.00667)	(0.0192)
Average tier		0.0104
		(0.00734)
Second lockdown × Average tier		-0.00590
		(0.00872)
Constant	-0.0449***	-0.0625**
	(0.0160)	(0.0258)
Day fixed effects	YES	YES
Observations	177074	177074

Note: This table refers to the impact of the second lockdown, vis-a-vis the same pre-pandemic period. Columns 1-2 feature effects for depression, while columns 3-4 feature effects for stress. Every specification accounts for time trends by including day fixed effects. Even columns include an interaction term for the average tier designation in the prior tiered lockdown period. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Impact of second national lockdown on mental health (cont.)

	(5)	(6)
	Anxious	
Days second lockdown	-0.000106	-0.000114
	(0.000429)	(0.000436)
Second lockdown	0.0191**	0.0367**
	(0.00869)	(0.0185)
Average tier		0.0114
		(0.0122)
Second lockdown × Average tier		-0.0104
		(0.00695)
Constant	-0.00353	-0.0228
	(0.00993)	(0.0283)
Day fixed effects	YES	YES
Observations	177074	177074

Note: This table refers to the impact of the second lockdown, vis-a-vis the same pre-pandemic period. Columns 5-6 feature effects for anxiety. Even columns include an interaction term for the average tier designation in the prior tiered lockdown period. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tables 17 to 21 present robustness checks for the main results. In particular, Table 17 considers the estimating equation in (7) with varying degree polynomials. Across the entire table, the results are very similar, which is suggestive of the fact that the results are not sensitive to the polynomial specification.

Table 18 verifies that the time trend is accounted for in the "second lockdown" period results. In particular, the estimation strategy assumes any variation in mental health states resulting from the progression of lockdown, conditional on the progression of days for the same pre-pandemic time period, is due to the lockdown per se. In other words, it assumes that time trends are the same across years, and that temporal variation in mental health trends is, loosely, seasonal. In order to test this, Table 18 performs a placebo test whereby equations (3) and (7) are estimated for a placebo period. It is held that if no effects are found for the comparison of 2018 to 2017 for the

Table 17: Robustness check: polynomial

	(1)	(2)	(3)	(4)
	Depressed			
Days second lockdown	0.000120	0.000119	0.000117	0.000117
	(0.000405)	(0.000405)	(0.000405)	(0.000405)
Second lockdown	0.00836	0.00835	0.00838	0.00838
	(0.00535)	(0.00535)	(0.00536)	(0.00536)
Constant	-0.0225**	-0.0225**	-0.0225**	-0.0225**
	(0.0101)	(0.0102)	(0.0102)	(0.0102)
Day fixed effects	YES	YES	YES	YES
Polynomial	5th	4th	3rd	2nd
Observations	177074	177074	177074	177074
	(5)	(6)	(7)	(8)
	Stressed			
Days Second lockdown	-0.000342	-0.000343	-0.000346	-0.000345
	(0.000340)	(0.000340)	(0.000339)	(0.000339)
Second lockdown	0.0368***	0.0368***	0.0368***	0.0368***
	(0.00667)	(0.00667)	(0.00666)	(0.00667)
Constant	-0.0449***	-0.0450***	-0.0449***	-0.0449***
	(0.0160)	(0.0160)	(0.0160)	(0.0160)
Day fixed effects	YES	YES	YES	YES
Polynomial	5th	4th	3rd	2nd
Observations	177074	177074	177074	177074
	(9)	(10)	(11)	(12)
	Anxious			
Days second lockdown	-0.000106	-0.000107	-0.000109	-0.000109
	(0.000429)	(0.000429)	(0.000429)	(0.000429)
Second lockdown	0.0191**	0.0191**	0.0191**	0.0191**
	(0.00869)	(0.00870)	(0.00870)	(0.00870)
Constant	-0.00353	-0.00357	-0.00355	-0.00355
	(0.00993)	(0.00993)	(0.00995)	(0.00995)
Day fixed effects	YES	YES	YES	YES
Polynomial	5th	4th	3rd	2nd
Observations	177074	177074	177074	177074

Note: This table refers to a robustness check that confirms that results in Tables 6 and 7 do not change according to the polynomial adopted in equation (1). Columns 1-4 refer to the second national lockdown's effects on depression, columns 5-8 refer to the effects on stress, and columns 9-12 refer to effects on anxiety. Columns 1, 5, and 9 feature a 5th order polynomial, columns 2, 6, and 10 feature a 4th order polynomial, columns 3, 7 and 11 refer to a 3rd order polynomial, and columns 4, 8, and 12 refer to a 5th order polynoomial. All specifications account for time trends by including a day fixed effect. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

second lockdown, then the time trends are adequately accounted for.[5] In sum, time trends seem adequately accounted for as there are no significant effects.

Table 18: Robustness check: time trends

	(1) Depressed	(2) Stressed	(3) Anxious
Placebo days second lockdown	-0.0000557	-0.000213	0.0000559
	(0.000292)	(0.000505)	(0.000425)
Placebo second lockdown	0.0109	0.00976	0.00865
	(0.00864)	(0.00865)	(0.00834)
	(0.0106)	(0.0174)	(0.0149)
Constant	-0.0315***	-0.0587***	-0.0226
	(0.0106)	(0.0174)	(0.0149)
Day fixed effects	YES	YES	YES
Observations	123079	123079	123079

Note: This table refers to a robustness check that confirms that results in Tables 6 and 7 are not driven by time trends. It performs the same analysis as the aforementioned tables, but for a placebo time period - namely the second national lockdown period, for 2018 versus 2017. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at the LAD level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tables 19 to 21 analyse whether the tiered lockdown results are due to the lockdown per se, or whether due to the overall pandemic. To this end, Table 19 features the tiered lockdown specification outlined in equation (6). In contrast to Table 14, it includes a control for monthly case rates at the Middle Layer Super Output Level. As is shown in this table, results persist, and in particular, Tier 2 designation led to an increase in anxiety, compared to Tier 1. However, Tier 3 designation no longer increases stress vis-a-vis Tier 1. This suggests that much of the stress caused by the pandemic was due to the case rates per se.

[5]The coronavirus epidemic started towards the end of 2019.

Table 19: Robustness check: impact of tiered lockdown on mental health (controlling for case rates)

	(1)	(2)
	Depressed	
Tier 2	-0.0299**	0.00267
	(-2.36)	(0.28)
Tier 3	-0.0169	0.0134
	(-1.22)	(1.05)
Case rate	0.000201	0.0000834
	(0.96)	(0.47)
User fixed effects	NO	YES
Observations	88118	88118
	(3)	(4)
	Stressed	
Tier 2	0.00267	0.000754
	(0.25)	(0.05)
Tier 3	0.0115	0.00173
	(1.19)	(0.12)
Case rate	0.000227	0.000351*
	(1.06)	(1.83)
User fixed effects	NO	YES
Observations	88118	88118
	(5)	(6)
	Anxious	
Tier 2	-0.00724	0.0158***
	(-0.78)	(3.31)
Tier 3	0.0128	0.0399***
	(1.22)	(2.72)
Case rate	0.000223	-0.0000272
	(1.51)	(-0.07)
User fixed effects	NO	YES
Observations	88118	88118

Note: This table refers to the impact of Tier 2 and Tier 3 (dummy variables) designation on mental health, with Tier 1 as the base category. Columns 1-2 refer to effects for depression, columns 3-4 refer to effects for stress, and columns 5-6 refer to effects for anxiety. Odd columns feature baseline OLS effects, while even columns feature user-level fixed effects. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tables 20 and 21 feature the specifications outlined in equations (7) and (9). In contrast to tables 15 and 16, however, these tables feature monthly case rates. As is shown in the tables, the results are largely robust to the inclusion of the case rates variable. Still, upon including the case rates variable, the effect of the second lockdown on depression gains significance.

Table 20: Robustness check: Impact of second national lockdown on mental health (controlling for case rates)

	(1)	(2)
	Depressed	
Days second lockdown	0.000443*	0.000441*
	(1.74)	(1.76)
Second lockdown	0.00521	0.0237
	(0.65)	(1.33)
Case rate	0.000131	0.000101
	(1.25)	(1.04)
Average tier		0.0140
		(1.15)
Second lockdown × Average tier		-0.00944
		(-0.99)
Constant	-0.0361***	-0.0631***
	(-4.97)	(-2.61)
Day fixed effects	YES	YES
Observations	322542	322542
	(3)	(4)
	Stressed	
Days second lockdown	0.000117	0.000126
	(0.46)	(0.49)
Second lockdown	0.0322***	0.0185
	(4.07)	(1.18)
Case rate	0.000286	0.000156
	(1.46)	(1.32)
Average tier		0.0121
		(1.19)
Second lockdown × Average tier		0.00747
		(0.93)
Constant	-0.0720***	-0.0948***
	(-10.17)	(-3.92)
Day fixed effects	YES	YES
Observations	322542	322542

Note: This table refers to the impact of the second lockdown, vis-a-vis the same pre-pandemic period. The first panel features effects for depression, while the second panel features effects for stress. Every specification accounts for time trends by including day fixed effects. Even columns include an interaction term for the average tier designation in the prior tiered lockdown period. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 21: Robustness check: Impact of second national lockdown on mental health (controlling for case rates)

	(1)	(2)
	Anxious	
Days seconds lockdown	-0.000293	-0.000287
	(-1.34)	(-1.31)
Second lockdown	0.0198***	
	(5.17)	
Case rates	0.000316*	0.000219
	(1.78)	(1.61)
Second lockdown		0.0153
		(1.53)
Average tier		0.0120
		(1.56)
Second lockdown × Average tier		0.00266
		(0.66)
Constant	-0.0198	-0.0425*
	(-1.55)	(-1.93)
Observations	322542	322542

Note: This table refers to the impact of the second lockdown on anxiety, vis-a-vis the same pre-pandemic period. Every specification accounts for time trends by including day fixed effects. Even columns include an interaction term for the average tier designation in the prior tiered lockdown period. All specifications include weighting to match SAMHI data, which is representative at the LSOA level. Standard errors in parentheses, and standard errors clustered at LAD level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.7 Conclusion

This study analysed the effect of coronavirus lockdown policies on mental health. Unlike previous studies that considered the mental health impacts of the pandemic, this study deployed quasi-experimental methods and Twitter data. This study also overcame representativeness challenges that were pending in use of Twitter data. With respect to the impacts of pandemic containment policies on mental health, this study found that, overall, lockdown policies had a negative effect on mental health. Furthermore, there were no potentiation effects whereby individuals in areas with more punitive tier systems experienced more mental ill-health in the second lockdown.

One of the key elements of this study - inferring mental health from large, high-frequency data, could be extended to contribute to the well-being economics agenda. In particular, while this study referred to mental health, a similar intuition could be used to make inferences about overall well-being.

Another extension of this work could be to understand whether the pandemic's mental health impacts had long-run effects. The windows of observation for this study are narrow, and could be expanded. Could there be a "long Covid" for mental health? Within this work, it would be interesting to understand which demographics are most affected by long-run pandemic mental health effects.

In a related vein, further work could understand whether the null potentiation effects found in this study mask heterogeneity. In particular, it is possible that individuals with prior better mental health did not suffer potentiation effects, while individuals with prior worse mental health did. Understanding these heterogeneous effects could be important to understanding mental health more generally.

Finally, this study contributed to better measurement of mental health. Future work could use large-scale data, such as Tweets, to answer causal questions about the roots of mental illness. Social scientists, with their training in quasi-experimental methods, are best-placed to answer questions about the social elements of mental illness. There is little causal evidence on things like the impact of childhood trauma or drug consumption on adult mental health. This work, and its data innovations, paves the way for answering these questions.

4 Valuing aesthetic features of the public realm: evidence from London town centres

4.1 Introduction

While public realm improvements - that is, the upgrading of the freely-accessible spaces between private buildings (Lofland (2017)) - are popular policies, little is known about their economic value. Partially, this is due to the fact that public realm improvements feature several different policy instruments. For instance, functional improvements that make the public realm more appealing to use (e.g. adding benches or creating spaces that make it easier for crowds to congregate), improving traffic flows, or purely aesthetic improvements. This study contributes to understanding how public realm improvements are economically valued by focusing on the last of these instruments - aesthetic improvements.

This study further breaks down the question of how aesthetic features of the public realm are valued, aiming to understand willingness to pay for facade architecture quality. Here, a revealed preference approach is taken, where capitalisation of facade quality into commercial rents proxies for the willingness to pay of passersby for facade quality. The basic intuition for this approach is that there is no explicit market for these environmental goods, such as canals for instance (Gibbons et al. (2019)). One way to obtain the value of these goods is to ask individuals (contingent valuation). Another strategy would be to look at how much individuals are willing to pay to be close to a particular environmental amenity. Holding all other housing factors constant, it is possible to infer marginal willingness to pay for an amenity by comparing house prices for homes that are closer/more distant to the amenity. In contrast to other work on the value of architecture (Ahlfeldt & Holman (2018)), this study considers how passersby (rather than home-owners) value facades. For this reason, we are interested in commercial rents, rather than house prices.

This study also contributes to the debate on the extent to which urban planning policies should protect distinctive architecture. While the potential pitfalls of urban planning regulation have been extensively studied, evidence defending the economic value of the design features these policies protect is scarce. Despite this lack of evidence, policy-makers tend to preserve architecture and design, often to the detriment of affordable housing. In fact, theory suggests that planning regulations increase land value (Hsieh & Moretti (2019)), which decreases affordable housing and office space and ultimately decreases urban productivity (Brueckner (1990)). Empirical evidence also shows that planning regulation increases housing price volatility (Hilber & Vermeulen (2016), Paciorek (2013)) and rent-seeking (Cheshire & Dericks (2014)). The consensus is that the costs of planning regulation outweigh benefits (Glaeser & Gyourko (2018)).

And yet, policy-makers continue to pursue regulation, particularly with respect to safe-guarding architecture and design. The rationale behind this persistence is that features of design are public goods, meaning that they will be under-provided without intervention and there may be important externalities to design, which spill over to adjacent plots and ultimately affect land prices. Whether or not this is a fair trade-off is still unknown, and there is a lack of evidence regarding the potential economic value architecture and design. Lack of evidence in this respect means that design can not be taken into account when performing cost-benefit analyses. This paper aims to help fill the gap in the literature, and follows Ahlfeldt & Holman (2018), who provide causal impacts of distinctive architecture on property prices.

One potential reason why the economic value of architecture is under-studied could be the paucity of data. Our paper contributes to data constraints inherent to this pending question in urban economics by using insights from data science. In particular, it leverages computer vision methods, which have made great advances in image feature extraction (see Krizhevsky et al. (2012) Girshick (2015), Chen et al. (2014)). Our approach makes use of Google StreetView images, and from there it extracts visual features from streets across London. We use a supervised learning approach, whereby images of facades were manually annotated according to their ornateness. It is worth noting that we consider "ornateness" because it is a tractable, non-loaded (by implicit valuations and notions of beauty), aesthetic feature. Consequently, we are agnostic about the direction of effects.

Another potential reason why the economic value of architecture is under-studied could also be due to the endogeneity of architecture with respect to economic outcomes. In our case, we're interested in the following question: if the quality of the facade design improves, all else equal, what effect might this have on commercial rents? The issue, in this particular case, is that e.g. more valuable parts of the city may attract distinctive buildings, and it is the geographic location of the buildings, rather than the architecture per se, that drives the commercial rents. To circumvent this, this study exploits the fact that British town centres are relatively homogeneous with respect to spatial characteristics, but, due to historical reasons, feature variation in facade quality. Accordingly, this study looks within London town centres, by using town centre fixed effects. Still, even within town centres there is some variation in geographic desirability, and for this reason, we use a town centre-level coordinate polynomial that accounts for within-town centre spatial trends.

Thee study finds that, once confounders are accounted for, streets with ornate facades that use materials such as glass and steel are signficantly more valuable in terms of commercial rents. This allows us to conclude, in a similar vein to Ahlfeldt & Holman (2018), that the planning debate should take into account the positive externalities inherent to good quality architecture.

The rest of the study describes in further detail the data. It then discusses the fixed effects strategy, followed by results. It concludes with some policy implications and potential avenues for future work.

4.2 Data

We collected three datasets for our study. The first dataset is Rateable Value data taken from the Valuation Office Agency (VOA) (VOA (2010)), which represents the annual rental value of a non-domestic property in an open market. This dataset provides our main independent variable of interest - value of commercial rents. Rateable values are fixed by an independent valuation officer from the VOA based on previous rental data and adjusted by the valuer to reflect the current condition of the property. While the precise calculations of rateable values are the subject of entire books, for the purposes of this study, it is essential to understand how the aesthetics of facades play into these calculations. This data could be problematic, if for instance, the surveyor implicitly attributes value to the aesthetics of facades, while these aesthetic factors do not in fact matter in real-world rent prices. Due to the way these rateable values are constructed, however, this is fortunately not a concern. Crucially, rateable values are calculated based on the actual rents of a previous year, and an appropriate time trend is added to this. This means that all "fixed effects" (including the facade) should be accounted for. In Figure 8, we plot rateable values against actual rents, for a sub-sample of properties for which we were able to extract Zoopla data. Note, a perfect 45 degree line would indicate a one-to-one mapping of rateable values to actual rents - both in annual British pounds. As the figure shows, the line is fairly close to this, and the R-squared shows that around 85 per cent of the variation in rateable values is explained by the actual rent amounts. It is worth noting that there are few observations due to the paucity of actual rent data.

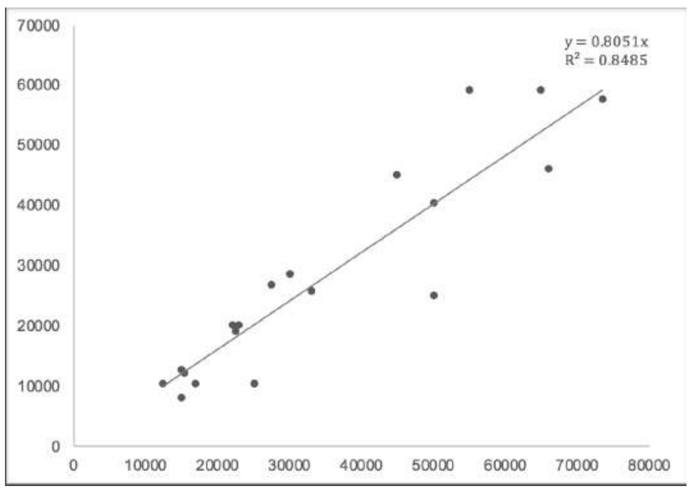


Figure 8: Rateable values versus actual rents

The second dataset is the London town centre boundaries CASA (2005). In particular, this dataset establishes an index of "town centredness", which includes things like retail density, and draws boundaries based on this index, for each postcode. The delineation of town centres is possible due to discontinuities in this index - for instance, the centroid of a town centre will have a high "town centredness" index, that becomes very low at the postcodes at the edges of this town centre. The dataset includes 256 town centres, such as Oxford Street (postcode W1) or High Street Kensington (postcode W8).

The third dataset is street images taken from the Google StreetView API Google (2018). Two building-facing images were collected for each street in the Greater London Area. We applying

Table 22: Ornateness Discussion Summary

Topics	Ornateness Questions
General	Is the facade functional or aesthetically crafted?
	Is the treatment of the facade significantly more decorated than necessary?
	Do the street facades exhibit flatness?
Facade	Do the street facades exhibit any change of materials or patterns (both between and within materials)
	Is the texture of the façade material intrinsic to the material itself?
	Does the choice of façade materials comply with the period in time?
Roof	Is the roofline of the façade consistent with structure of the roof?
	Is its cornice highly decorative beyond is function?
Doors	Does door have features other than a basic frame?
	Is the shape of the door regular?
Windows	Does the window-hood have feature other than basic horizontal beams?
	Is the shape of the window regular?

a supervised machine learning approach, whereby we trained the machine learning model with a series of annotated images. We classified each image on a binary regarding whether the facade was ornate or not. We followed a series of rules, that are outlined in Table 20. At the core of our approach was attributing a score of 1 to facades for which any element of the architecture was unnecessarily decorative. For instance, with respect to the "Does building exhibit any change of materials or patterns (both between and within materials)" question, it is held that varied materials are suggestive of ornamentation.

Given that the Streetview images are at the "street" level, it is worth defining precisely what is meant by "street" level. Henceforth, a 'street" refers to a street segments, where each segment is defined as the edges between junctions. For instance, Oxford street would have many segments. Figure 9 provides an illustration of this.

The fourth dataset that is used also street images taken from the Google StreetView API Google (2018). It refers to the period of materials used in building facades. Here, streets are classified according to whether a majority of the facade in the image uses old or new materials. New materials are glass and steel, while older materials include brick.

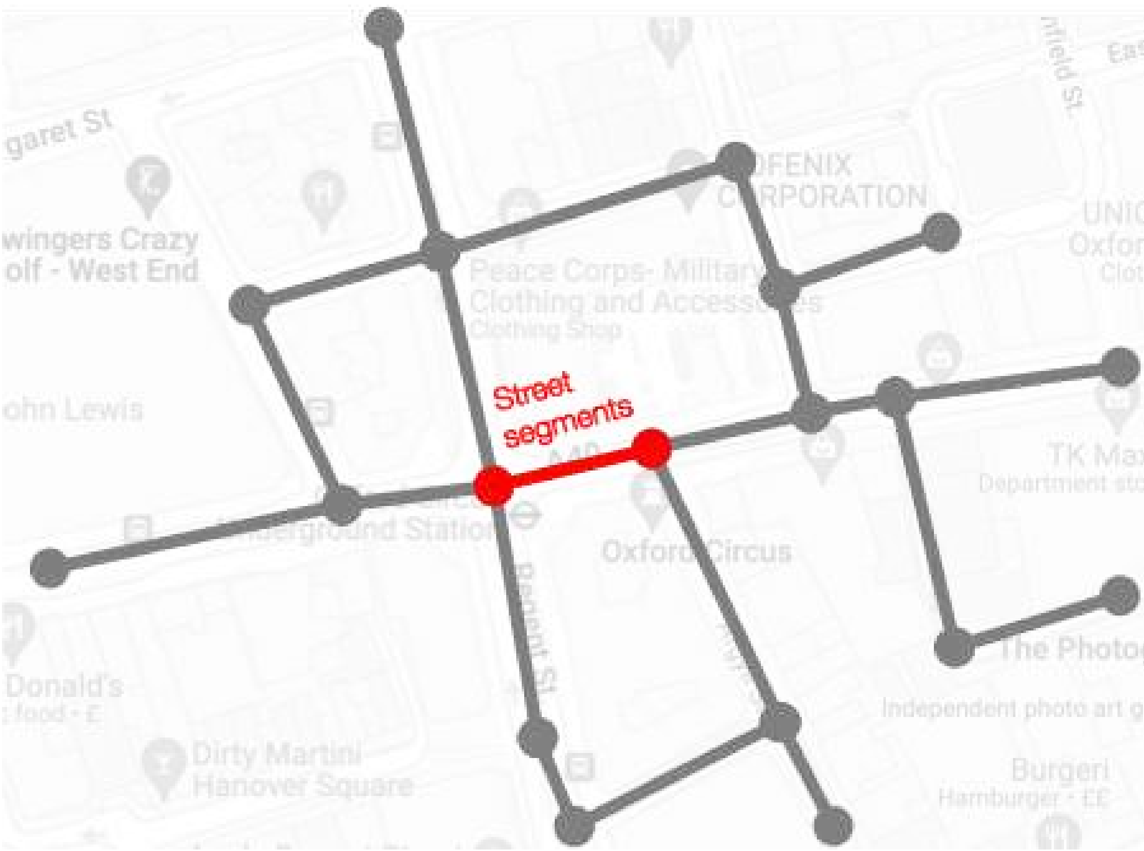


Figure 9: Street segnment

To summarise, our dataset was constructed using several sources. The main independent variable is rateable value, which serves as a credible proxy for commercial rents. This variable is available the street level. The independent variable of interest - architectural ornateness - is available at the "street" level. Here, a "street" includes both sides of the street.

The explanatory variable of interest is a dummy that takes the value of one if either side of the street is ornate. In other words, we consider the capitalisation effects of the architecture of the entire street, of which own-facades are a subset.

Finally, our control variables include building-level characteristics such as floor area, precise location coordinates, and geographic accessibility.The metric of geographic accessibility is measured as a sum of jobs e divided by distance d within 60 minutes, $g_i = \sum e_j d_{ij}^{-1}$ where i is each retail unit and j is each Lower Super Output Area (LSOA). The distances in this metric are travel time, excluding national rail, and it is assumed that travelling via the street is 5km/h and tube/dlr/overground is 15km/h. This variable tells us the extent to which a particular building is connected to job poles in the rest of the city, and controls for transport links that may be driving the building's commercial rent. A limitation of this variable is that it excludes national rail, and therefore introduces measurement error for a series of locations that are accessible in this way. Furthermore, we use a variable that distinguishes whether town centre features predominantly retail/hospitality commercial units or office space.

Table 23 features the descriptive statistics for the aforementioned variables.

Table 23: Descriptive statistics

	Mean	SD	Observations
Rateable value	59324	236366	38826
Ornateness	0.657	0.474	38826
New materials	0.40	0.49	38826
Floor area	208	1232	38826
Geographic accessibility	35211	38232	38826

4.3 Estimation Strategy

In order to understand whether individuals value aesthetic features of the public realm, in an ideal scenario we would have exogenous variation in architectural facade quality. In this scenario, the baseline equation would amount to:

$$Y_{bsc} = \alpha + \beta_1 Ornate_s + \beta_2 X_b + \epsilon_{bsc} \tag{10}$$

Where Y_{bsc} is the rateable value for building b in street s in town centre c; $Ornate_s$ is a dummy that takes the value of 1 if the street is ornate on any side, 0 if ornate on no sides; X_b is a vector of building-specific controls, including floor area.

In this scenario, even with exogenous ornateness, it is necessary to control for the interior aspects of the building, as we are only interested in aspects of the public realm - namely, facades. Unfortunately, ornateness is not exogenous, and simple correlation (conditional on controls) of ornateness and commercial rents is likely upwardly biased. For instance, it can be held that certain architectural characteristics are correlated with location, and it is this location that is driving the commercial rents, rather than the quality of the public realm per se. In order to illustrate this idea, it is worth outlining some of London's architectural history. London can be thought of as a patchwork of independent towns that were, with time, subsumed into the central city limit. As Jim Clifford writes in A Mighty Capital Under Threat: The Environmental History of London, 1800-2000 (p.23-24):

> Beginning in the late eighteenth century, regions on London's urban fringe generally developed first into market gardens or dumping sites, with ribbons of development spreading along the main roads. Gradually, the land between the thoroughfares filled in with suburban streets and row houses, or a cluster of factories. This process repeated over time as new regions on the outskirts formed into suburbs, and the older districts became inner suburbs or a part of London. Hackney, for example, transformed from a rural hinterland in 1800 to a growing suburb in the second half of the nineteenth century, and it was increasingly absorbed into London proper during the twentieth

> century. Preexisting towns, villages, and hamlets also transformed into suburbs as London's sphere of influence and transportation networks expanded. This resulted in satellite suburbs growing in Outer London, beyond the limits of the continuous built-up area. Croydon is the best early example; it started as an independent town south of London in Surrey, but increasingly became a major commuter suburb during the nineteenth century.

Two features of London, which will be useful for our estimation strategy, are implicit in this quote. The first feature is that London can be thought of as a poly-centric city (Kloosterman & Musterd (2001)), where independent towns were subsumed into the city limits. Crucially, these towns retain some of their independence, in the sense that each town has a centre where most of its commerce is centralised. Within these town centres, things like planning regulations, local clientele, and surrounding house prices are held constant.

However, the fact that these things are held constant within town centres is not necessarily useful for estimation, unless there is within-town centre variation in architecture. The above quote argues that these town centres, which were absorbed by the inner city, were not developed at once, and instead, developed gradually. This means that a given town centre may feature the style of architecture that was en vogue at the time that the rural parcel was converted to urban use. In fact, there are striking differences in English architecture through the ages (Atkinson (1928)).

In order to understand whether there is sufficient variation to perform this analysis, it is necessary to outline the standard deviations of ornateness. Across the whole sample, the standard deviation is 0.474. The average standard deviation of within-town centre ornateness, on the other hand, is 0.159. While this may seem low relative to the entire-sample standard deviation, these within-town centre standard deviations range from 0 to 0.976. This follows anecdotal evidence on the discrepancies in town centre sizes.

These two features of London architecture allow us to look within town centres and thereby control for a series of important confounders. Even within town centres, however, there might be persistent correlations between location and public realm features. One example might be that buildings along Oxford Street change non-randomly as one walks from Tottenham Court Road to

Marble Arch. It is the e.g. proximity to Hyde Park that drives these effects, rather than the style of the buildings, even within Oxford Street. A solution to this is to include a polynomial function of location, which accounts for spatial trends that are correlated with design, and in turn affect commercial rents.

In this case, it is likely that a local (at the level of the town centre) polynomial will more adequately account for spatial trends than a global (at the whole sample level) polynomial. To illustrate, the parameter of a eastings coordinate, if calculated globally, can not adequately account for the fact that e.g. converging east in one part of the city implies a rise in commercial rents, while converging east in another part of the city implies a decrease in commercial rents. To account for the unique trends, it may be preferable to calculate the polynomial at the town centre level. In practice, one way to achieve this is to interact the town centre dummies with the coordinates. However, even though it is in principle preferable to have a local location function, there may be a trade-off between polynomial order and granularity of the polynomial. Given that we are using two variables (i.e. coordinates) in our polynomial, and that these variables and their interactions must be multiplied by town centre dummies, using a polynomial higher than second order is impossible given the current sample size.

The analysis below, in Table 24, seeks to understand whether it is preferable to use a higher-order global polynomial, or a second-order local polynomial. In particular, it performs a town centre fixed effects regression of rateable values on the location polynomial, and records the within-town centre R-squared. The goal is to understand which specification of the polynomial explains more variation in the rateable values data. The first line of Table 22 features the within R-squared of the global, second-order polynomial. Only 0.03 percent of the variation of rateable values is explained by this location function. Adding global polynomial parameters to account for third-degree trends, as noted in line two of Table 22, only increases the accounted variation by 0.02 percentage points. Finally, line three includes the within R-squared second-order town centre-level, which accounts for 1.5 percent of the variation in rateable values. Clearly, the local, but lesser-order polynomial is preferable to a more flexible, global polynomial.

Table 24: R-squared for different polynomials

	Within R-squared
Global 2nd order	0.0003
Global 3rd order	0.0005
Local 2nd order	0.015

Note: This table features the within town centre R-squared for a regression of rateable values on town centre fixed effects and a given polynomial function. The first two lines feature the within R-squared for a polynomial function defined globally (i.e. the coordinate variables are not interacted with a town centre dummy). The third line features the within R-squared for a polynomial function defined locally (i.e. the coordinate variables are interacted with a town centre dummy).

With the order of polynomial and level at which parameters should vary, the preferred specification's estimating equation amounts to:

$$Y_{bsc} = \omega_c + \beta_1 Ornate_s + \beta_2 X_s + \beta_3 f_c(x_b, y_b) + \epsilon_{bsc} \tag{11}$$

Where ω_c is a town centre fixed effect; Y_{bsc} is the rateable value for building b in street s in town centre c; $Ornate_s$ is a dummy that takes the value of 1 if the street is ornate on any side, 0 if ornate on no sides; X_b is a vector of building-specific controls; $f_c(x_b,\, y_b)$ is function, at the town centre level, of the geographic coordinates.

It is worth noting that in spite of controlling for certain building-specific characteristics, it is possible that ornate street facades are correlated with other building characteristics. For instance, facade design is often correlated with interior styles of architecture. It is possible that e.g. glass buildings let in more light, and it is this light that is felt on the inside of the building that is driving the results, rather than the glass facade per se. To account for these factors, we consider heterogeneous effects by average building material, at the street level. In particular, we distinguish between two types of building materials: old and new. Old building materials include brick, while new building materials include glass and steel. Accordingly, our final specification is:

$$Y_{bsc} = \omega_c + \beta_1 Ornate_b \times Material_b + \beta_2 Ornate_b + \beta_3 Material_b + \beta_4 X_s + \beta_5 f_c(x_b, y_b) + \epsilon_{bsc} \quad (12)$$

4.4 Results

Table 25 features the OLS effects of street ornateness on rateable values or commercial rents for retail and hospitality units. Column 1 does not include any controls for building characteristics, while column 2 does. In particular, without controlling for building characteristics, having ornateness at the street level does not seem to affect rental values.

It is possible that the ornateness variable in column 1 is explained by other correlated aesthetic features. For instance, ornate streets may use older or more modern materials, which tend to be more or less aesthetically pleasing. To understand whether this is true, the next columns interact ornateness with these potential confounders. Columns 3 and 4 feature the OLS effects of ornateness, interacted with whether the street featured new materials, on commercial rents. This "new materials" variable takes the value of one if the majority of the street featured materials such as glass and steel, and zero if the majority of the street featured materials like brick. The odd column features these effects without building-level controls, while the even column features these effects with building controls. Without building controls, there does not seem to be an effect of ornateness on retail/hospitality commercial rents. However, column 4 shows that the interaction term is significant. In particular, buildings in streets that use more modern materials such as glass and steel and are also ornate pay a £46596 annual premium, compared to buildings in non-ornate streets that use old materials. The ornateness variable is significantly negative, which means that buildings in streets that use old materials and are ornate pay £25371 less in annual rents than old-material, non-ornate buildings. Finally, the new materials variable is also negative, which suggests that buildings in non-ornate that use new materials also pay £28343 less in annual rent than buildings in old-material, non-ornate streets.

Table 25: OLS effects of ornateness on retail/hospitality commercial rents

	(1)	(2)	(3)	(4)
	Rateable value			
Ornateness	-4066.3	-506.6	-23773.9	-25371.0***
	(10541.6)	(4334.3)	(16850.5)	(6956.3)
Floor area		185.1***		185.1***
		(0.868)		(0.866)
Geographic accessibility		1.105***		1.153***
		(0.0863)		(0.0869)
New materials street			-15672.7	-28343.3***
			(19323.1)	(7997.9)
New materials street × Ornateness			42856.4	46596.4***
			(22121.3)	(9151.9)
Constant	126247.4***	-27366.9**	136466.5***	-13275.6
	(9206.5)	(8894.7)	(15603.2)	(9827.4)
Town centre fixed effects	NO	NO	NO	NO
Geographic control function	NO	NO	NO	NO

Note: This table features effects for the baseline OLS specification of ornateness (dummy variable) on rateable values. Column 1 features the OLS specification without control variables, column 2 features the OLS specification with control variables, column 3 features the OLS specification without control variables but with an interaction term for material age (a dummy variable that takes the value of one if materials are glass or steel), column 4 features the OLS specification with control variables and the interaction term. Floor area and geographic accessibility are calculated at the building level, while all other variables are calculated at the street level. Standard errors in parentheses, and standard errors clustered at town centre level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

However, the coefficients in Table 25 may be biased by the the fact that certain town centres, located in more/less desirable parts of the city, may have more ornate buildings. Furthermore, within town centres, certain sections may be located in more desirable areas, and may also feature more ornate buildings. Accordingly, Table 26 controls for these sources of endogeneity, by using town centre level fixed effects and a town centre-specific coordinate polynomial control function. Both columns feature building-level controls. The first column shows no effect on commercial rents. However, as before, it is possible that ornateness is actually serving as a proxy for certain architectural materials. Column 2 features ornateness interacted with whether the buildings in the street used new materials, such as steel or glass. Compared to buildings in non-ornate streets that use old materials, buildings in ornate and new-material streets command a £52095 annual rent premium. As in the OLS specification, the ornateness coefficient is negative, which suggests that buildings in ornate and old-material streets pay £33832 less in rent than buildings in non-ornate and old-material streets. Similarly, the new materials coefficient is negative, which suggests that buildings in non-ornate and new-material buildings pay £30082 less in rent than buildings in non-ornate and old-material streets.

Table 26: Town centre FEs effects of ornateness on retail/hospitality commercial rents

	(1)	(2)
	Rateable value	
Ornateness	-5500.9	-33832.0**
	(4658.2)	(7602.5)
Floor area	184.5***	184.4***
	(0.840)	(0.839)
Geographic accessibility	3.876***	3.900***
	(0.310)	(0.310)
New materials street		-30082.7*
		(8521.2)
New materials street × Ornateness		52095.3*
		(9648.2)
Town centre fixed effects	YES	YES
Geographic control function	YES	YES
Observations	9315	9315

Note: This table features effects for the town centre fixed effects specification of ornateness (dummy variable) on rateable values. Column 1 features the fixed effects specification with control variables, column 2 features the fixed effects specification with control variables and with an interaction term for material age (a dummy variable that takes the value of one if materials are glass or steel). All specifications include the geographic control function, which is a local (at town centre level) second order polynomial of coordinates.Floor area and geographic accessibility are calculated at the building level, while all other variables are calculated at the street level. Standard errors in parentheses, and standard errors clustered at town centre level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To summarise, ornate, glass or steel buildings are more valuable than non-ornate brick buildings. In contrast, ornate brick buildings are less valuable than non-ornate brick buildings, as are non-ornate glass or steel buildings. Put differently, if a building is made with new materials, it pays to be ornate. On the other hand, if a building is made with old materials, it pays to be less ornate. Table 25 summarises these results, with cells presenting the average prices of buildings within the categories.

These effects can be thought of as the valuation of different architectural trends. In order to better understand the results, it is worth mapping the four material age-ornateness categories to broad architectural styles. In "Ornament and Crime" (Loos (2019)), Adolf Loos writes a manifesto for Modernist architecture. This style was notable for its use of glass and steel, and favouring non-ornate and strictly functional facades (Rowe (2011)). In response the modernist movement's austerity, post-modernist architecture featured highly ornate facades, and used a range of materials (Clendinning (2002)).

Mapping our results to recent architectural styles, our analysis shows that individuals tend to value buildings modernist buildings the most, after buildings that use new materials but are ornate. Old material buldings that are non-ornate (e.g. terraced houses) are valued the second-least, while broadly postmodern buildings are valued the least.

4.5 Robustness check

As a robustness check, a placebo analysis is performed. This placebo test entails looking at the main effects specified in equation (2) for commercial units we would not expect to see an effect for - office space. The intuition for this placebo test is that if facades are capitalised into commercial rents, this must be driven by the fact that environmental quality (i.e. facades) and products sold in commercial spaces (e.g. coffee) are complements in consumers' utility functions. The fact that

Table 27: Summary of main results

	Old materials	New materials
Ornate	-£33832	+£52095
Non-ornate	Base category	-£30082

we find effects for retail/hospitality buildings suggests that environmental quality increases the demand for the goods therein sold. We would not expect to see such results for office spaces, as these do not sell goods for which there are potential complementarities, and if the results for office spaces are significant, then the results in Table 28 are driven by something other than facades per se. Indeed, the main interaction term (new materials × ornateness) does not have a significant coefficient. In other words, the results in Table 26 are driven by the facades themselves. It is worth noting, however, that a limitation of this placebo test may be that larger service-oriented firms may indeed be willing to pay more for ornate facades, as this would add credibility to the firm's brand. The fact that the results are null, however, seems to imply that these firms do not appreciably weigh in the sample.

Table 28: Effects of ornateness on office rents (placebo)

	(1)	(2)
	Rateable value	
Ornateness	-5027.2	-6257.4*
	(1205.9)	(1654.6)
Floor area	103.6***	103.6***
	(0.540)	(0.540)
Geographic accessibility	1.223*	1.191*
	(0.270)	(0.270)
New materials street		-4017.7
		(1817.0)
New materials street × Ornateness		1335.1
		(2357.4)
Town centre fixed effects	YES	YES
Geographic control function	YES	YES
Observations	29511	29511

Note: This table features effects for the town centre fixed effects specification of ornateness (dummy variable) on rateable values, for office rents. This table provides a placebo test for the effects in Table 5. All specifications include the geographic control function, which is a local (at town centre level) second order polynomial of coordinates. Floor area and geographic accessibility are calculated at the building level, while all other variables are calculated at the street level. Standard errors in parentheses, and standard errors clustered at town centre level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.6 Conclusion

This study used a novel approach to infer architectural features from Google Streetview images. It then applied a town centre fixed effects and location polynomial to uncover the effect of facade ornateness on commercial rents. It contributes to literature on the economic value of architectural design, ultimately contributing to the debate on whether planning regulation in the UK is excessive. Furthermore, it builds on scarce evidence regarding public realm investments. It finds that architectural features indeed capitalised into commercial rents, and that buildings in ornate streets that use newer materials are particularly valuable.

Given that this study fits within two research streams, there are two potential avenues for further work. The first entails further contributing to the planning debate literature. In particular, this study has analysed the extent to which facades are capitalised into commercial rents, and ultimately, whether individuals are willing to pay for aesthetically pleasing public realm. Finding that individuals are indeed willing to pay for these features, aesthetics should be introduced into future cost-benefit analyses of planning regulation. Other work could focus on the positive externalities that come from aesthetically-pleasing facades that may not be capitalised into commercial rents. Such benefits could relate to wellbeing.

The second stream of research that could benefit from further work is the public realm investment impact evaluation literature. In particular, analysing the impact of the Future High Street Funds, which aimed to rejuvenate high streets in the UK, could be a fruitful starting point.

More generally, while we have shown that aesthetically pleasing facades are indeed economically valuable, still little is known about whether beautification policies lead to commercial or even residential displacement. Finally, work should take into account the larger question of whether it is possible to both provide higher quality public realm, without causing displacement.

4.7 Appendix: Streetview data technical summary

This section will describe the methods that allowed us to infer architectural characteristics (our main explanatory variable) from StreetView images. Several recent works provided evidence that street-level images of a city can be used to estimate socio-economic differences.

Naik et al. (2014) collected human perception data from street images through a crowd-sourced survey (Place Pulse 2.0). Dubey et al. then used this data Dubey et al. (2016) by fitting a multi-stream convolutional neural network model (CNN) to predict perceived safety, which are likely important covariates in a retail rental value model. Seresinhe et al. (2017), similarly, collected perceived scenicness data from a crowd-source survey and found evidence that urban scenicness is indeed related to urban wellbeing and reported health outcomes. Gebru et al. (2017) in contrast extracted highly interpretable features such as car types from Google StreetView images to correlate with socio-economic factors such as income, voting and demographic patterns in the States. The authors found that these interpretable medium-level features can be used to predict the income, race, education, and voting patterns at both the zip code and precinct level for cities in the States. Different from previous work, Law et al Law et al. (2018) used street and satellite imagery to estimate a visual response from house prices. Our research extends this work, but rather than learning a visual response that estimates house price, we learn here a mapping between a sparse set of visual components and ornateness.

Methodologically, our pipeline consists of three stages. First, we annotate the image as illustrated in 4.7.1. Second, we trained a supervised model as illustrated in 4.7.3, to classify whether a street facade image are ornate or not ornate using the annotated data. Third, we use this trained model to infer ornateness for every single street in London and translate this into a binary variable for the estimation.

In particular, if the image in question had one of the characteristics as illustrated in 4.7.3, it was considered "ornate". At the crux of all the specific architectural characteristics that we identify as ornateness is the simple notion of purely aesthetic embellishments versus functionality. In other words, if a facade featured embellishments that served no purpose - i.e. did not contribute to structural integrity of the building - then it was considered "ornate". In sum, following this rule of thumb allowed us to be sensitive to context, ultimately providing us with a way to discern whether certain features were "ornate" or not. An example of feature ambiguity could be a column, which, in certain architectural styles, serves no purpose other than ornamentation. In others, it is

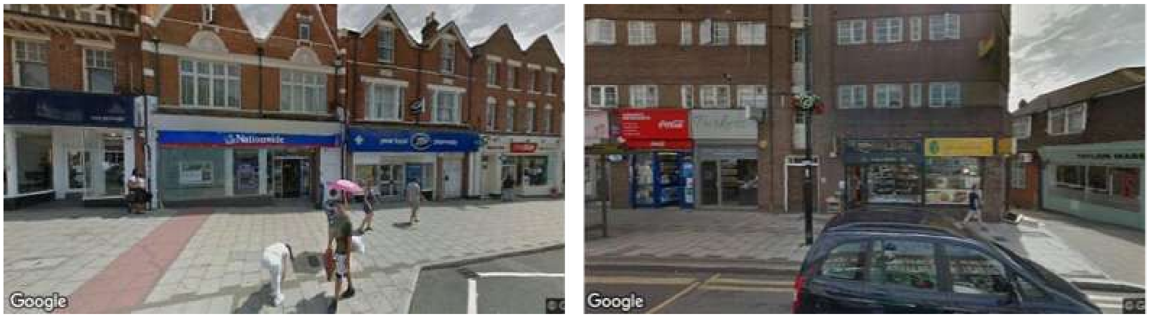


Figure 10: **Left:** Ornate building facade **Right:** less Ornate building facade

important for the building's structural integrity.

4.7.1 Ornateness: ground-truth annotation

This section will describe the data annotation process for the *OrnateCNN* model. Ornateness is defined where the majority of the building facade has more ornamentation than those constructed in a similar time period. We use this definition to ensure we reduce the biases where older are often perceived to be more ornamental than newer ones. The process is as of follows;

- Sample 2000 building facade images in London using the Google StreetView API. For each sampled image;
- Filter whether the facade is valid (perpendicular street facade) or not valid (not perpendicular, blank image, interior image, obstructed image)
- Classify whether both stories of the building is ornate or not

This process had been conducted by the two authors (an urban designer/urban analytics researcher and an economic geographer). The process includes an initial phrase where each person annotated 100 street facade images with meta-description followed by a discussion. A summary of the discussion is consolidated into a set of questions to ask when we annotate the building facade image. We then split the dataset into two halves and annotated the samples as the ground-truth building ornateness data.

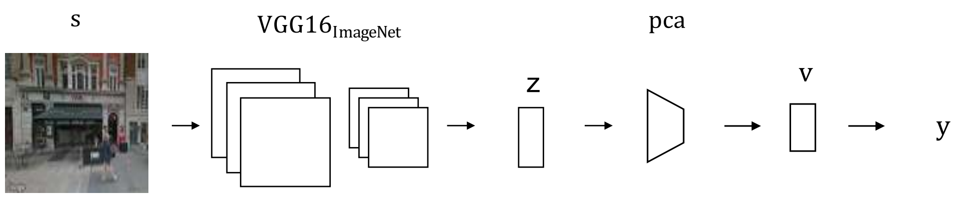


Figure 11: the architecture of OrnateCNN

4.7.2 Ornateness CNN Model

We propose here the OrnateCNN model ($OrnateCNN$) which combines a standard pretrained Convolutional Neural Network architecture (CNN) with a linear PCA (PCA_{lin}) to estimate the ornateness of building facades in London using the data collected from the previous section.

In the first step we used a pretrained Convolutional Neural Network CNN to extract visual latent features z from the images. The feature extractor is a pretrained $VGG16$ CNN model Simonyan & Zisserman (2014) learnt from the ImageNet dataset Deng et al. (2009) where the last fully connected layer is dropped. The feature vector z has a dimension of 4,096.

In the second step, we apply a linear principal component analysis PCA_{lin} that summarises the visual feature z into a set of linearly uncorrelated and ordered components v. To compute PCA, we first standardised the data and compute the Eigenvectors and Eigenvalues of the feature covariance matrix. We then take the first 64 principal components (leading Eigenvectors) v as visual summaries for each building facade.

In the third and final step, we train a logistic regression model that maps the visual summaries v to the ornateness y (0,1). The parameters of the logistic regressor $y = F_w(v)$ are updated by minimising the following binary cross entropy loss function where y_i is the ground truth ornateness label and p_i is the predicted probability of whether the facade is ornate or not.

$$L_r = -\frac{1}{n} \sum y_i log(p_i) + (1 - y_i) log(1 - p_i) \tag{13}$$

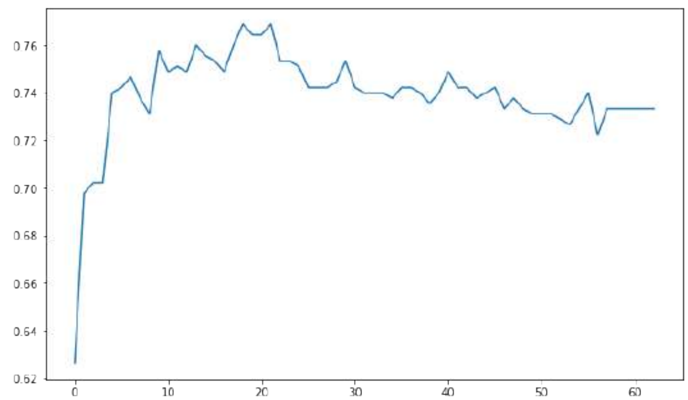


Figure 12: Out-of-sample ornateness accuracy plot. We estimated a set of logistic regressor and report here the accuracy in the y-axis and the number of components to include in the x-axis. We achieve an optimal out-of-sample accuracy of 78% with 19 principal components.

4.7.3 Ornateness Experiment

To identify a sparse set of component for the *OrnateCNN* model, we conducted the following ornateness experiment. We first split the ground truth dataset into a train and test set (75% and 25%). We then estimate a set of logistic regression model sequentially where we vary the number of components v from $(1, 64)$. The results of the ornateness experiment is illustrated in fig 12 where we achieve an out-of-sample test-set accuracy of 78% between the ground truth and the predicted labels with only 19 principal components. The predicted ornateness are visualised in 13 where we observe higher levels of ornateness in Central London relative to outer London. Specifically the areas south of the Thames and Canary Wharf have lower values of predicted ornateness.

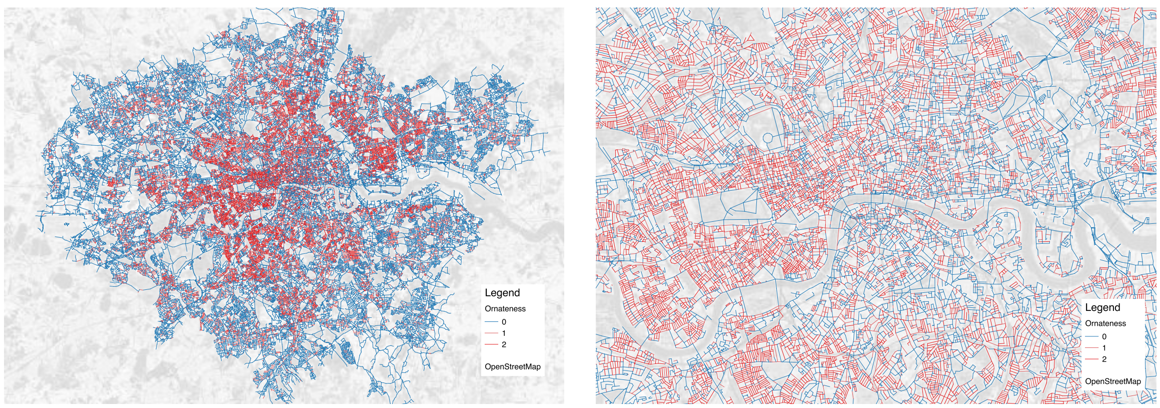


Figure 13: **Left:** London Building Ornateness/Richness. Red denotes higher visual ornateness/richness and blue denotes lower visual ornateness/richness. **Right:** A zoomed in plot of the building Ornateness/Richness in Central London. Contains Ordinance Survey data ©Crown copyright and database right ©2018.

5 Comparing causal estimates of neighborhood effects on parenting behavior

5.1 Introduction

This study considers whether the parenting practices of different types of families - namely families who live in public versus private housing - are more or less sensitive to neighborhood crime.

The first set of findings comes from the Moving to Opportunity (MTO) experiment which randomly allocated housing vouchers to families residing in low-income public housing in five U.S. cities. Comparing households across voucher randomization groups, Leventhal & Brooks-Gunn (2005) have found small effects of neighborhoods on several measures of parenting behavior. We replicate these findings.

In turn, we compare the above results to those obtained from a different sample - namely the Project for Human Development in Chicago Neighborhoods (PHDCN). This study surveyed a group of families living in Chicago, outside of public housing. In the absence of a randomized experiment, causal estimates of neighborhood effects are identified using a quasi-experimental research design that relies on the impact of administrative public housing closures on other Chicago neighborhoods. Han (2021) finds relatively larger and statistically significant neighborhood effects on parenting behavior.

We study the difference in findings by constructing parenting behavior scales that are identically measured in both samples. We create harmonized index measures of parental warmth, harshness, and monitoring behavior.[6] Even when using these harmonized measures, the discrepancy in results persists. In other words, the different findings of neighborhood effects cannot be explained by differences in measurement of outcomes.

We explore two potential explanations for the discrepancy in neighborhood effect estimates across samples. First, we consider differences in neighborhood impacts driving the changes in parenting behavior in each sample. The MTO intervention was explicitly targeted at reducing

[6] While we retain the names of these measures from Leventhal & Brooks-Gunn (2005), the measures we create are not necessarily identical to theirs. The exact construction of these measures is not described, and we have excluded certain behavior variables that do not appear identically in both samples.

neighborhood poverty in the treated population. In contrast, the administrative closures of public housing caused untargeted changes on different aspects of neighborhoods for the PHDCN sample. To assess whether differences in neighborhood impact is a likely explanation, we restrict the MTO sample to households who were likely to use the voucher to move to lower crime neighborhoods. Since the PHDCN sample experienced measurable changes to neighborhood crime, this restriction on the MTO sample results in greater comparability between samples in neighborhood attributes affected. We also consider whether differences in household characteristics is a likely explanation for the discrepancy in findings. To assess this explanation, we reweight the MTO sample to closely resemble the PHDCN sample in several demographic characteristics. In both cases, we find that the differences between MTO and PHDCN neighborhood effects persist. This means that the above two explanations are insufficient to explain the discrepancy in findings. We conclude the paper by discussing some likely explanations that are not ruled out by our analysis.

5.2 Theoretical framework

This section provides a literature-based, theoretical framework for our study. It provides an overview of the neighborhood effects literature, and makes the case for differential neighborhood effects for public versus private housing families.

This literature review departs from the fact that, taken at face value, the quasi-experimental literature on neighborhood effects is contradictory. It then synthesises the literature by making generalisations within categories of studies. It argues that there is a fair amount of consensus in the literature, and in particular, that peers matter depending on the granularity of the geography in question. Importantly, it makes the case the correlated amenities matter depending on the populations in question - i.e. social housing and non-social housing populations. Ultimately, it argues that there are two possible explanations for discrepancies between populations. The first is the possibility of complementarities between neighborhood and family investments, in a similar vein to the early education literature. The second is that the same variables that determine selection into social housing also interact with neighborhood effects on parenting. This framework provides a justification for our empirical exercise, in which we compare the impact of neighborhoods on parenting for different populations.

A close dissection of the experimental neighborhood effects literature suggests that findings are actually aligned with each other. To expand upon this, it is worth first outlining the seemingly disparate nature of effects. With respect to the impact of neighbourhoods on jobs, some studies find no effects (Oreopoulos (2003)), while others find significant effects (Chetty & Hendren (2018), Cutler & Glaeser (1997), Ananat (2011), Edin et al. (2003), Damm & Dustmann (2014), (B¨""ohlmark & Willén (2020), Kondo & Shoji (2019), Chyn (2018)). With respect to welfare receipt, Åslund & Fredriksson (2009) find that neighborhood peers matter. Similarly, with respect to university attendance, Barrios Fernández (2019) finds that close neighbors have an affect on whether someone attends university. Results are again more divided with respect to test scores, with Gibbons et al. (2013), Weinhardt (2014), and Del Bello et al. (2015) finding that neighborhoods have no effect on test scores. On the other hand, Goux & Maurin (2007) finds that close neighbours are an important determinant of test scores. Finally, one study finds that neighborhoods are relevant determinants of criminal behavior (Damm & Dustmann (2014)).

This apparent lack of consensus is partially driven by the fact that there are two types of quasi-experimental studies. The first considers the impact of neighbourhood peers per se. The second considers the effect of peers, as well as all other correlated effects, which include amenities and other location characteristics.

Looking within these categories, it is possible to make some generalisations. Studies that isolate peer effects per se tend to not find any effects when large geographic units are considered (Weinhardt (2014), Gibbons et al. (2013)). However, evidence points to the fact that close neighbors matter (Barrios Fernández (2019), Goux & Maurin (2007)). Furthermore, Rotger & Galster (2019), who identify peer effects from the quasi-random influx of criminal neighbors, find that neighborhood peer effects matter for crime. An exception to these results is Del Bello et al. (2015), who conclude that school peers matter more for test outcomes than neighborhood peers.

Evaluations of the Moving to Opportunity policy fall within the category of “correlated effects” studies. This is because the experiment changed not only the neighborhood peers, but also, all the amenities that are a result of low-poverty demographics. Studies found that the program improved adult mental health (Katz et al. (2001), Leventhal & Brooks-Gunn (2003), Sanbonmatsu et al. (2011), Kling et al. (2007)). Furthermore, Leventhal & Brooks-Gunn (2003) find that mental health also improved for boys but not girls, likely because girls are more sheltered with respect to neighborhood effects. With respect to economic self-sufficiency and jobs, studies find no effects (Katz et al. (2001), Sanbonmatsu et al. (2011), Kling et al. (2004)).

In short, to summarise a vast literature on MTO impacts, the program had no effect on most outcomes except for adult mental health. However, recent MTO evidence finds that moving to a low-poverty neighbourhood at a young age has positive effects for long-term employment and educational outcomes (Chetty et al. (2016)). On the other hand, quasi-experimental studies that do not disentangle peer effects per se from correlated effects tend to find that neighborhoods do matter. One exception to this is Oreopoulos (2003), who finds that families who were quasi-randomly allocated better neighborhood fared no better than those allocated worse neighborhoods.

Clearly, there is a puzzling discrepancy between experimental and quasi-experimental studies within the "correlated" effects neighborhoods literature. It is worth, then, categorising these studies further in an attempt to make generalisations. As noted, MTO studies find that neighborhoods in general don't matter for economic self-sufficiency. Oreopoulos (2003) echoes these conclusions. What these two sets of studies have in common is the population they consider. Namely, families residing in social housing. These findings are suggestive of heterogeneous effects of neighbourhoods by family income. However, it is difficult to make these inferences as contexts vary wildly between studies. Our study fills this gap, by conducting a like-for-like comparison of neighborhood effects on parenting, within the same context, for social housing and non-social housing populations.

As will be shown, we find discrepancies in effects between these two populations. Two hypotheses justify these differences in effects. The first hypothesis is that there are dynamic complementarities between family investments and neighborhoods. The second is that the same unobservable variables that determine selection into social housing then interact with neighborhood investments to yield parenting practices.

With respect to the first hypothesis, the basic intuition is that social housing determines family investments, and these in turn interact with neighborhood investments to yield differences in parenting practices. Indeed, one reading of Chetty et al. (2016) corroborates this idea. Children who spent most their lives in better neighborhoods were able to accrue the advantages that come from, e.g., better amenities. On the hand, older children were not. In other words, there might be complementarities between neighborhood investments and other early educational investments that are typically made by families.

However, Chyn (2018), in a study similar to Oreopoulos (2003), find that children of social housing families who moved to better neighborhoods experienced better outcomes. At face value, this contradicts what was previously outlined. However, the authors note that in contrast to Oreopoulos (2003), their context entailed a much larger change in neighborhood context. To synthesise and extrapolate, these findings don't contradict the idea of complementarities, but rather, suggest that effects can be found provided the neighbourhood investments are sufficiently large.

Parallels can be drawn between this idea and the dynamic complementarities literature in education. This literature argues that education investments, particularly early education investments, interact with future investments (Cunha & Heckman (2007)). In particular, recent causal evidence shows that K-12 investments have higher yields for low-income families if interacted with early education investments (Johnson & Jackson (2019)). On the other hand, Malamud et al. (2016) find that family investments do not interact with educational investments due to compensatory behavior.

Dynamic complementarities are hard to prove. One way of thinking about dynamic complementarities is to borrow from the mediation literature (see e.g. Heckman & Pinto (2015) for a review of this literature). In other words, the dynamic complementarities problem can be reformulated as: to what extent are the causal impacts of investments in X on Y mediated by investments in Z? As noted within the mediation literature (Frölich & Huber (2014), and further reiterated by Johnson & Jackson (2019), answering these types of questions requires two sources of exogenous variation, in both X and Z, and are therefore challenging to tackle.

This paper does not aim to answer definitively whether such complementarities exist for neighborhoods, families, and parenting practices. This is mostly because while we have exogenous variation in neighborhood quality, we do not have exogenous variation in family social housing status. Instead, our findings may be suggestive of the second hypothesis, which holds that the same unobservable characteristics that determine family social housing status in turn interact with neighborhoods to yield different parenting practices.

To reiterate, while we do not isolate dynamic complementarities or unobservable selection effects, we make the case that both of these might be relevant. This paves the way for future work.

5.3 Neighborhoods and parenting literature

This section provides an overview of the neighbourhoods and parenting literature. The impact of neighbourhoods on family processes is well-studied in the literature, although very few studies move beyond correlation. This literature review will describe studies according to the family processes they consider. The first group of studies considers the impact of neighborhoods on parental monitoring. Parental monitoring includes, for instance, parental knowing of child whereabouts, and knowledge regarding peer groups. These studies can be further divided into which aspects of neighborhoods they consider. With respect to neighborhood social cohesion (that is, the quality of social ties between neighbors), Zuberi (2016) and Byrnes & Miller (2012) find that better social cohesion leads to higher parental monitoring, and knowledge of child whereabouts specifically. With respect to neighborhood SES more generally, results are mixed. In particular, Chung & Steinberg (2006) finds no effect of neighborhood SES on monitoring, while Zuberi (2016) and Dishion & McMahon (1998) find negative effects, and Chuang et al. (2005) find positive effects. On the other hand, Byrnes et al. (2011) find that higher SES neighborhoods are associated with more rules but less knowledge of whereabouts. These contradictory results within the broader category of "monitoring" may provide justify the rather inconclusive results in other studies. Byrnes & Miller (2012) also consider the impact of neighborhood social cohesion on parenting styles, and find that socially supportive neighborhoods lead to less authoritarian parenting. Other studies consider the impact of neighborhoods on parental involvement more broadly defined. They find either null (Elliott et al. (2016)) or negative effects associations between neighborhood deprivation and parental involvement. Greenman et al. (2011) find that low SES neighborhoods are associated with less monitoring of child homework. Finally, another set of studies consider the effect of neighborhoods on parental warmth and nurturing. The most statistically robust of these studies finds that higher SES neighbourhoods increase parental harshness towards female children (Leventhal & Brooks-Gunn (2005)). Similarly, a less robust study (Brody et al. (2001)) finds that community disadvantage is associated with less nurturing parenting, but parental nurture in bad neighbourhoods made the most difference. On the other hand, Tendulkar et al. (2010) finds that neighborhood poverty has no effect on parental warmth. Other studies consider a range of different neighborhood characteristics. Tendulkar et al. (2010) find that neighborhood social support and neighborhood safe places to play had no effect on parental warmth. McDonell (2007) find that neighborhood policing increased parental warmth. With the exception of Leventhal & Brooks-Gunn (2005), who leverage MTO data, all the aforementioned studies are purely correlational.

5.4 Background & Data

This section describes the two main policies and data sources that were used. These two sources, as noted, will provide different effects that will be compared. This section will also outline how we harmonize variables between the two data sources.

5.4.1 The Moving to Opportunity (MTO) Intervention

The Moving to Opportunity intervention was implemented by the Department of Housing and Urban Development (HUD) in five US cities: Baltimore, Boston, Chicago, Los Angeles, and New York. From September 1994 to July 1998, families living in public and assisted housing were randomly assigned to three groups: two voucher treatment groups and a control. The Section 8 group received the standard rent-assistance voucher (of the same name). This voucher was eligible for use in all neighborhoods.[7] The experimental group received a voucher that could only be used in neighbourhoods with a poverty rate of ten per cent or lower, i.e. fewer than ten percent of the families earned below the federal poverty line. They also received counselling regarding how to find a private rental unit. The voucher amount for both groups the gross rent for the unit minus

[7]The terms of the Section 8 voucher are as follows: families pay 30 per cent of their income towards rent. The remainder will be paid by the government, up to a rent cap of the Fair Market Rent for that city.

30 % of monthly adjusted income. These families were then re-surveyed 4 to 7 years after initial randomization (interim evaluation), and also 10 to 15 years after (final evaluation). To ensure a time frame of analysis that is comparable to the PHDCN, we use outcome variables measured at the time of the interim evaluation.

5.4.2 The Project for Human Development in Chicago Neighborhoods (PHDCN) Study

The Project for Human Development in Chicago Neighborhoods was a comprehensive study of the pathways of human development in children and youth. To this end, a sample of children and adolescents were initially surveyed between 1994 and 1996. Subsequently, follow-up surveys were conducted in 1996-1998 (Wave 2) and 2000-2002 (Wave 3).

In order to measure the impact of family, neighborhood and school environment, the study recorded key features of each social context. In particular, parenting behavior was measured at each wave, so that repeat measurements are available for the same parent.

5.4.3 Comparing Variables

In order to make the estimates as comparable as possible, we harmonize the main outcomes of interest, as well as the neighborhood attribute used to evaluate neighborhoods. The main outcomes of interest are parental warmth and harshness. Following Tendulkar et al. (2010) and Leventhal & Brooks-Gunn (2005), we define each outcome as an index score aggregating interviewer observations on the presence/absence of a particular kind of parenting behavior during the course of the interview.[8] Crucial to our study, the parenting behavior items comprising each score are identical in the MTO and the PHDCN data.

One key problem with both sets of data is measurement error. These variables may not reflect the day-to-day behaviour of parents, as they may have modified their attitudes towards their children in light of the interviewer's presence. This is problematic if certain types of parents chose to

[8] For parental warmth, parenting behavior items include "Does parent use a positive tone of voice when speaking to the child?" and "Does parent answer child's questions?". For parental harshness, items include "Did parent shout at child?" and "Did parent respond to the child in an annoyed manner?"

be less revealing of their true practices. On the other hand, there may be variation in the interviewers' interpretations of parenting behaviour. This is problematic if the MTO interviewers were systematically different to the PHDCN interviewers.

Due to data confidentiality requirements, the definition for a neighborhood is slightly different in the MTO and PHDCN samples. In the MTO, a neighborhood is defined by the census tract. For the PHDCN, a neighborhood is defined using a study-specific unit of the neighborhood cluster: an aggregation of relatively similar census tracts. Hence, a PHDCN neighborhood is slightly larger than a neighborhood cluster is roughly 2.5 census tracts in size.

Table 29: Neighborhood Attributes, MTO Sample vs PHDCN Sample

	Low-Poverty	Traditional	Control	Chicago	PHDCN	All Chicago
Proportion of White Residents	0.29	0.27	0.23	0.11	0.40	0.43
	(0.27)	(0.25)	(0.23)	(0.18)	(0.28)	(0.37)
Poverty Rate	0.32	0.34	0.42	0.41	0.22	0.23
	(0.18)	(0.15)	(0.15)	(0.18)	(0.11)	(0.17)
Crime Rate	0.12	0.096	0.10	0.098	0.085	0.11
	(0.51)	(0.41)	(0.31)	(0.026)	(0.039)	(0.07)
Observations	1500	1100	1200	750	2411	

Note: Table shows means and standard deviations (in parentheses) of neighborhood attributes experienced by MTO and PHDCN sample. MTO sample is divided into by randomization group - "low poverty", "traditional", and "control". Separately, statistics for the Chicago sub-sample of MTO (all 3 randomization groups) are presented. The definition of a neighborhood for the MTO is the Census tract, and for the PHDCN the neighborhood cluster (aggregation of roughly 2.5 Census tracts, designed specifically for the survey). "All Chicago" refers to every Census tract in Chicago.

We characterize neighborhoods using neighborhood crime rate. Here, the crime rate is the number of criminal ocurrences per capita. This is mainly for practical reasons: which can be reliably matched to neighborhoods in both samples. Since the focus of our analysis is cross-sample comparisons, we make no attempt to isolate the impact of crime from the impact of other neighborhood variables. Therefore, the coefficient estimates should be interpreted as the impact of crime along with other correlated neighborhood characteristics. For the purposes of comparing neighborhood quality effects, it is not a primary concern if other neighborhood attributes were changing along with crime rate (as is most likely the case). However, it is a more serious concern if the MTO intervention changed a different set of neighborhood attributes relative to the PHDCN. While we are limited in our capacity to address related concerns, we perform some sensitivity analysis to check the robustness of our results.

5.5 Methodology

The empirical analysis involves the between-sample comparison of regression-based estimates of residential neighborhood effects on parenting behavior. Because the MTO is experimental in nature while the PHDCN is not, we employ different empirical strategies to obtain the relevant estimates for each sample. We describe each of these in turn.

$$y_i = \alpha^M 1\left(v_i \in \{s8, MTO\}\right) + X_i^{M\prime}\beta^M + \varepsilon_i^M \tag{14}$$

For the MTO sample, we follow existing research in estimating intent-to-treat effects of housing voucher receipt, following regression equation (14). y_i denotes one of the selected parenting behavior outcomes for household i and v_i denotes the voucher group that the household was randomized into. Hence, the coefficient α^M measures the impact of being offered a housing voucher, as opposed to actual use of the voucher (which is a choice made by the household). [9] Following the literature, we also control for X_i^M: a set of household covariates measured during the baseline survey, as well as city fixed effects. These controls are included to improve the precision of the estimates and identification α^M does not rely on their inclusion.

For the PHDCN sample, we follow the quasi-experimental approach used in Han (2021) in order to estimate the effect of neighborhood crime on parenting. The regression equation is described in (15). The parenting variables are now additionally indexed by time t, since repeated measures per household are available in the PHDCN. $c_{j,t}$ denotes the crime rate in the household's chosen neighborhood j and in year t.

$$y_{i,t} = \alpha^P c_{j,t} + \varepsilon_{i,t}^P \tag{15}$$

In the absence of experimental randomization, we instrument for neighborhood crime using

[9] The results are qualitatively unchanged if we use separate indicators for each voucher group, since each voucher group has a similar impact on parenting as well as on neighborhood crime rate.

the predicted impact of administrative public housing closures. While PHDCN households were not residing in public housing, they were nonetheless affected by the inflow of displaced residents. Hence, the quasi-experimental variation in neighborhood quality arises from the spillover impact of public housing closures and not the direct impact. Our measure of the predicted impact of housing closures is described in (16): we use the product of 1990 neighborhood black share (b) with the yearly number of housing units closed (h) in all other neighborhoods. Here, we use public housing closures in other neighborhoods due to the fact that families in the PHDCN sample generally do not live in public housing neighborhoods. This measure captures variation in the yearly level of closures, as well as spatial variation in the impacts of the closures, which is strongly correlated with the existing racial composition of the neighborhood.

As an additional feature to eliminate potential bias due to neighborhood selection, we first estimate (16) using a yearly panel of all neighborhoods in Chicago. Aside from the instrument, we control for fixed effects at the neighborhood level as well as a quadratic time trend. After generating predicted crime rates for each neighborhood based on the estimated coefficients, we assign values of predicted crime rate to households based on the first observed neighborhood of the household, ignoring any subsequent residential moves.

$$c_{j,t} = \delta b_{j,1990} \sum_{k \neq j} h_{k,t-1} + \gamma_1 t + \gamma_2 t^2 + \eta_{j,t} \tag{16}$$

To summarize our instrument, it is the extent of public housing closures in adjacent neighbourhoods, interacted with the share of own-neighbourhood black residents. Thus, the instrument forcibly captures the effects of the influx of black residents into the neighbourhood, following the demolition of nearby social housing projects. The rationale for this instrument is that residents of social housing projects are typically black, and MTO evidence suggests that individuals tend to choose neighbourhoods with predominantly own-race residents. In sum, the instrument captures an exogenous change in neighbourhood demographic composition, and this demographic change may also shift crime levels.

In this study, we only require the instrument to satisfy the following exclusion restriction:

changes in unobserved household characteristics across time must be uncorrelated with the 1990 black share of the neighborhood. This exclusion restriction does not preclude all forms of neighborhood sorting by PHDCN families, which is highly improbable. Because (15) includes a family-level fixed effect, any sorting based on fixed characteristics does not bias the results. Instead, the exclusion restriction only requires that, conditional on the family's initial situation, there was no further sorting based on (anticipated) future changes to unobserved characteristics.[10]

However, there are other threats to the exclusion restriction. Firstly, it is possible that black share interacted with social housing closures affects parenting practices through changing the social fabric of the neighbourhood. For instance, it is possible that a predominantly black neighbourhood attracts more social housing residents, and given that they knew each other before, they are better able to share caring costs for their children. The final result would be for parents to reduce their monitoring.

$$c_j = \alpha^C 1\left(v_i \in \{s8, MTO\}\right) + X_i^{M\prime} \beta^M + \varepsilon_i^C \tag{17}$$

At this point, the MTO coefficient (α^M) and the PHDCN coefficient (α^P) correspond to different magnitudes of change in neighborhood crime. To make the PHDCN estimates comparable to the MTO estimates, we first estimate the impact of housing voucher receipt on neighborhood crime using the MTO sample. We estimate (17), which is identical to (16) except that the dependent variable is now the neighborhood crime rate. Next, we multiply the PHDCN coefficient (α^M) by the estimated value of α^C. We compare this scaled neighborhood effect ($\alpha^M \alpha^C$) with the MTO coefficient, since both now represent the same change in neighborhood crime rate.

We evaluate several likely explanations for any differences that we find. One such explanation is different sample demographic composition. We evaluate this explanation by reweighting the MTO sample to resemble the PHDCN sample in terms of primary caregiver age, race and education. Using the PHDCN sample, we estimate sample weights for each age $\times$ race $\times$ education cell of the primary caregiver.[11] We then reweight the MTO sample using these estimated weights and

[10] Han (2021) presents some evidence against this kind of sorting, showing that changes in observable characteristics between survey waves appear uncorrelated with the initial neighborhood Black share.

[11] Age is coded into 4 categories ("Under 35", "36 to 40", "41 to 45" "46 and over"), education is coded as a

re-estimate α^M. If demographic differences are the main driver of different neighborhood effects, this reweighting procedure should yield MTO and PHDCN coefficients that are very similar.

Another potential explanation for different neighborhood effect estimates is that the impact of experimental voucher randomization differed from the quasi-experimental impact of public housing closures, in terms of neighborhood characteristics other than crime. While we measure neighborhoods using the crime rate, it is highly likely that the identifying variation in each sample affected other neighborhood characteristics as well.

5.6 Results

Table 30 presents our comparison of neighborhood effects between the MTO sample and the PHDCN sample. Starting with the MTO estimates, we present estimates of the impact of receiving either type of housing voucher. Column 1 shows that the effect on parental warmth is negative insignificant - statistically as well as in terms of the magnitude of the point estimate. To demonstrate the latter finding, we present the mean and standard deviation of parental warmth among the control group. According to the point estimate, the impact of voucher receipt on parental warmth is less than 5 percent of the control group standard deviation. Column 2 shows that the impact of voucher receipt on parental harshness is similarly insignificant. In this case, the point estimate is an order of magnitude smaller relative to the control group standard deviation. Taken together, these two columns show little evidence of MTO intervention impacts on parenting behavior specifically.

To reconcile our findings with existing research, we use a dependent variable that is more comparable to ones used in prior studies: an index of parental monitoring. In column 3, we find that housing voucher receipt has a positive and statistically significant impact on parental monitoring, with a magnitude slightly over a tenth of a standard deviation for the control group. We reiterate that this variable measures the extent to which parents know about their child's whereabouts and companions. Therefore, one potential explanation for the isolated voucher impact on parental monitoring is that the child's behavior is being altered instead of the parent's. It is

binary variable for having a high school degree, and race is coded into 3 categories ("White/Other", "Black" and "Hispanic").

Table 30: Neighborhood Effects on Parenting Behavior, MTO Sample vs PHDCN Sample

	(1)	(2)	(3)	(4)	(5)	(6)
			MTO			PHDCN
	Warmth	Harshness	Monitoring	Crime	Warmth	Harshness
Intent-to-treat: Any Voucher	-0.0915	0.00783	0.0462	-0.0598	-1.200	-0.190
	(0.177)	(0.0423)	(0.0201)	(0.0241)	(0.008)	(0.127)
Control group mean	6.289	0.268	0.440	-2.797		
Control group SD	2.306	0.684	0.378	0.755		
Observations	950	1300	2000	2200	7129	7129

Note: This table features the impact of crime on parenting for MTO and PHDCN . Columns 1-4 feature outcomes for MTO, while columns 5-6 feature outcomes for PHDCN. For the MTO sample, the coefficients show intent-to-treat impact of receiving either Section 8 or Experimental voucher. For the PHDCN sample, the coefficients show instrumental variable estimate for impact of neighborhood crime, rescaled to match MTO estimates. The column headers indicate dependent variable. The Warmth/Harshness/Monitoring variables are and index measure for corresponding parenting behavior. The Crime variable features the number of crimes per resident (logarithmic scale).

possible that the voucher intervention resulted in more predictable routines and/or social networks for the child. As a consequence, parents would report more accurate knowledge about their child's routines without any increase in monitoring effort.

We present scaled neighborhood effect estimates from the PHDCN in Columns 5 and 6. Column 5 shows a negative and statistically significant reduction in parental warmth. Column 6 also shows a reduction in parental harshness, although the point estimate falls short of statistical significance. For both parental warmth and parental harshness, the PHDCN estimates of neighborhood effects are noticeably larger than the MTO coefficients. Overall, it appears that neighborhood effects are sizable for the PHDCN sample but negligible for the MTO sample.

We evaluate two likely explanations for this discrepancy. In Table 31, we consider differences in sample demographics as a potential explanation. After reweighting the MTO sample by age/education/race of the household head to match the PHDCN sample, we find that the effects of voucher receipt are essentially unchanged. The scaled neighborhood effects from the PHDCN are slightly reduced because of the smaller MTO voucher impact on neighborhood crime. However, the magnitudes of the PHDCN estimates still remain much larger than the MTO estimates, suggesting that sample demographics are not the main explanation.

We also consider differences in the neighborhood characteristics affected. Because the MTO

Table 31: Neighborhood Effects on Parenting Behavior - MTO Sample Reweighted to Resemble PHDCN

	(1)	(2)	(3)	(4)	(5)	(6)
			MTO			PHDCN
	Warmth	Harshness	Monitoring	Crime	Warmth	Harshness
Intent-to-treat: Any Voucher	-0.118	0.0437	0.0694	-0.0413	-0.829	-0.131
	(0.197)	(0.0461)	(0.0229)	(0.0257)	(0.475)	(0.0874)
Control group mean	6.277	0.243	0.436	-2.734		
Control group SD	2.3	0.625	0.373	0.717		
Observations	950	1300	2000	2200	7129	7129

Note: This table features the impact of crime on parenting for MTO and PHDCN samples, with the MTO reweighted to match PHDCN on the following variables: age, education, and race. Columns 1-4 feature outcomes for MTO, while columns 5-6 feature outcomes for PHDCN. For the MTO sample, the coefficients show intent-to-treat impact of receiving either Section 8 or Experimental voucher. For the PHDCN sample, the coefficients show instrumental variable estimate for impact of neighborhood crime, rescaled to match MTO estimates. The column headers indicate dependent variable. The Warmth/Harshness/Monitoring variables are and index measure for corresponding parenting behavior. The Crime variable features the number of crimes per resident (logarithmic scale).

intervention incentivized choosing neighborhoods with lower poverty and not neighborhood crime, MTO voucher recipients experienced changes in neighborhood crime that were quite non-uniform. To test whether households experiencing reductions in neighborhood crime were more responsive, we restrict the MTO sample to households that were likely to experience reductions in neighborhood crime. We assess this likelihood by regressing neighborhood crime changes on household baseline characteristics, restricting the sample to the Section 8 voucher group. We then generate predicted values of neighborhood crime for the entire MTO sample, and re-estimate our MTO estimates for the subsample with predicted reductions in crime.

Table 32 shows the results of this analysis. On this subsample, the voucher impact is larger than our baseline results, but still small relative to the control group standard deviation and the PHDCN estimates. Therefore, it appears that the varied impact of MTO on neighborhood crime is not the explanation for the discrepancy in neighborhood effects.

Table 32: Neighborhood Effects on Parenting Behavior - MTO Sample Predisposed Towards Lower Crime Neighborhoods

	(1)	(2)	(3)	(4)	(5)	(6)
	MTO				PHDCN	
	Warmth	Harshness	Monitoring	Crime	Warmth	Harshness
Intent-to-treat: Any Voucher	-0.219	-0.0313	0.0373	-0.0898	-1.80	-0.285
	(0.204)	(0.0497)	(0.0233)	(0.0265)	(1.03)	(0.190)
Control group mean	6.109	0.256	0.421	-2.748		
Control group SD	2.406	0.674	0.375	0.76		
Observations	750	1000	1500	1700	7129	7129

Note: This table features the impact of crime on parenting for MTO and PHDCN samples, restricting the MTO sample to families that chose low-crime neighbourhoods. Columns 1-4 feature outcomes for MTO, while columns 5-6 feature outcomes for PHDCN. For the MTO sample, the coefficients show intent-to-treat impact of receiving either Section 8 or Experimental voucher. For the PHDCN sample, the coefficients show instrumental variable estimate for impact of neighborhood crime, rescaled to match MTO estimates. The column headers indicate dependent variable. The Warmth/Harshness/Monitoring variables are and index measure for corresponding parenting behavior. The Crime variable features the number of crimes per resident (logarithmic scale).

5.7 Explanations for discrepancies

This section outlines the limitations of our analysis, and in particular, potential reasons underlying the discrepancies between our cross-sample coefficients.

The first issue relates to the extent to which our demographic matching works. In particular, we rely on a limited set of demographic variables to match our coefficients. It is possible that some important differences that determine selection into social housing are driving our results. However, as noted in the theoretical framework, this study does not aim to identify dynamic complementarities between social housing status and neighborhood investment. In sum, this shortcoming is in fact a part of the research question we are aiming to answer, as we are partially interested in whether selection into social housing interacts with neighborhood investment.

Concerns that detract from our research question, however, are also relevant. One key concern is that quasi-experimental variation in crime is identified in different ways across samples. In the MTO sample, families were forced to move to different neighborhoods. In the PCDHN sample, changes in crime are identified via the influx of new residents - i.e. these families did not move. It is possible that the effects of moving to a new neighborhood are driving the differences in results.

Further work could address this by understanding whether MTO effects are in part driven by the displacement per se. There were also important income effects related to the MTO voucher.

Related to these issues is also the concern of correlated neighborhood amenities. In short, while in the PCDHN sample only crime and related amenities changed, in the MTO sample, all amenities related to neighborhood poverty changed. Further work could address these concerns by better controlling for these correlated amenities.

Finally, there are underlying methodological issues that make comparing MTO to PCDHN difficult. In particular, while both RCT (MTO) and IV (PCHDN) rely on a sub-population of compliers to identify effects, these compliers may not be comparable across samples. That is, for each population (social housing and non-social housing), compliers may not be representative. This means that any effects that are uncovered may not be telling of differences between populations.

5.8 Conclusion

In this paper, we document differences between causal estimates of neighborhood effects: comparing the results of an experimental housing intervention (MTO) and quasi-experimental variation in neighborhood crime caused by public housing closures (PHDCN). Using the same measures of parenting behavior, we find that neighborhood effect estimates are small and insignificant in the MTO sample. In contrast, with the PHDCN sample, we find that the corresponding change in neighborhood crime is associated with a large decrease in both parental warmth and harshness, which is statistically significant in the case of parental warmth. Through reweighting, we rule out two potential explanations for the different estimates: differences in sample demographics, as well as differing magnitudes of change in neighborhood crime.

This analysis allows us to validate what a speculative reading of the neighborhood effects literature suggests. Namely, that neighborhoods affect families who opt into social housing (at any point in past) differently. Through a like-for-like comparison, we conclude that this is the case.

As noted, however, this study makes no claims regarding what this ultimately implies. There are two possible explanations for these discrepancies. The first is aligned with the dynamic complementarities literature in education. It suggests that there are interactions between family investments in social housing settings and neighborhood investments. Another possibility is that the same factors that lead a family to self-select into social housing interact with the effects of neighborhoods on parenting. While this study does not disentangle these two forces, it does pave the way for future research in this vein.

References

Aaronson, D., Hartley, D. & Mazumder, B. (2017), The effects of the 1930s holc ””redlining”” maps, Technical report, Working Paper.

Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. M. (2022), ‘When should you adjust standard errors for clustering?’, *The Quarterly Journal of Economics* **138**(1), 1–35.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. (2011), Sentiment analysis of twitter data, *in* ‘Proceedings of the Workshop on Language in Social Media (LSM 2011)’, pp. 30–38.

Ahlfeldt, G. M. & Holman, N. (2018), ‘Distinctively different: a new approach to valuing architectural amenities’, *The Economic Journal* **128**(608), 1–33.

Aisopos, F., Papadakis, G. & Varvarigou, T. (2011), Sentiment analysis of social media content using n-gram graphs, *in* ‘Proceedings of the 3rd ACM SIGMM international workshop on Social media’, ACM, pp. 9–14.

Aloufi, S. & El Saddik, A. (2018), ‘Sentiment identification in football-specific tweets’, *IEEE Access* **6**, 78609–78621.

An, X., Ganguly, A. R., Fang, Y., Scyphers, S. B., Hunter, A. M. & Dy, J. G. (2014), Tracking climate change opinions from twitter data, *in* ‘Workshop on Data Science for Social Good’.

Ananat, E. O. (2011), ‘The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality’, *American Economic Journal: Applied Economics* **3**(2), 34–66.

Aono, M. & Himeno, S. (2018), Kde-affect at semeval-2018 task 1: Estimation of affects in tweet by using convolutional neural network for n-gram, *in* ‘Proceedings of The 12th International Workshop on Semantic Evaluation’, pp. 156–161.

Asiaee T, A., Tepper, M., Banerjee, A. & Sapiro, G. (2012), If you are happy and you know it... tweet, *in* ‘Proceedings of the 21st ACM International Conference on Information and Knowledge Management’, ACM, pp. 1602–1606.

Asiamah, N., Opuni, F. F., Mends-Brew, E., Mensah, S. W., Mensah, H. K. & Quansah, F. (2021), 'Short-term changes in behaviors resulting from covid-19-related social isolation and their influences on mental health in ghana', *Community Mental Health Journal* **57**(1), 79–92.

Åslund, O. & Fredriksson, P. (2009), 'Peer effects in welfare dependence quasi-experimental evidence', *Journal of Human Resources* **44**(3), 798–825.

Aston, N., Liddle, J. & Hu, W. (2014), 'Twitter sentiment in data streams with perceptron', *Journal of Computer and Communications* **2**(03), 11.

Atkinson, T. D. (1928), *English Architecture*, .

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M. & Varma, V. (2012), Mining sentiments from tweets, *in* 'Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis', pp. 11–18.

Barbieri, F. & Saggion, H. (2014), Modelling irony in twitter, *in* 'Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics', pp. 56–64.

Barbosa, L. & Feng, J. (2010), Robust sentiment detection on twitter from biased and noisy data, *in* 'Proceedings of the 23rd international conference on computational linguistics: posters', Association for Computational Linguistics, pp. 36–44.

Barrios Fernández, A. (2019), 'Should i stay or should i go? neighbors' effects on university enrollment', *Neighbors' effects on university enrollment (May 1, 2019)* .

Becker, G. S. (1973), 'A theory of marriage: Part i', *Journal of Political economy* **81**(4), 813–846.

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M. & Kane, J. M. (2017), 'A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals', *Journal of Medical Internet Research* **19**(8), e289.

Black, S. E. (1999), 'Do better schools matter? parental valuation of elementary education', *The Quarterly Journal of Economics* **114**(2), 577–599.

B¨""ohlmark, A. & Willén, A. (2020), 'Tipping and the effects of segregation', *American Economic Journal: Applied Economics* **12**(1), 318–47.

Bora, N. N. (2012), 'Summarizing public opinions in tweets', *International Journal of Computational Linguistics and Applications* **3**(1), 41–55.

Bosco, C., Patti, V. & Bolioli, A. (2013), 'Developing corpora for sentiment analysis: The case of irony and senti-tut', *IEEE Intelligent Systems* **28**(2), 55–63.

Bravo-Marquez, F., Mendoza, M. & Poblete, B. (2013), Combining strengths, emotions and polarities for boosting twitter sentiment analysis, *in* 'Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining', ACM, p. 2.

Brody, G. H., Conger, R., Gibbons, F. X., Ge, X., McBride Murry, V., Gerrard, M. & Simons, R. L. (2001), 'The influence of neighborhood disadvantage, collective socialization, and parenting on african american children's affiliation with deviant peers', *Child Development* **72**(4), 1231–1246.

Brueckner, J. K. (1990), 'Growth controls and land values in an open city', *Land Economics* **66**(3), 237–248.

Burnap, P., Colombo, W. & Scourfield, J. (2015), Machine classification and analysis of suicide-related communication on twitter, *in* 'Proceedings of the 26th ACM Conference on Hypertext & Social Media', ACM, pp. 75–84.

Byrnes, H. F. & Miller, B. A. (2012), 'The relationship between neighborhood characteristics and effective parenting behaviors: The role of social support', *Journal of Family Issues* **33**(12), 1658–1687.

Byrnes, H. F., Miller, B. A., Chen, M.-J. & Grube, J. W. (2011), 'The roles of mothers' neighborhood perceptions and specific monitoring strategies in youths' problem behavior', *Journal of youth and adolescence* **40**(3), 347–360.

CASA (2005), http://www.casa.ucl.ac.uk/towncentres/cd/tcmap.htm.

Caspi, A., Bolger, N. & Eckenrode, J. (1987), 'Linking person and context in the daily stress process.', *Journal of Personality and Social Psychology* **52**(1), 184.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. (2014), 'Semantic image segmentatiom with deep convolutional nets and fully connected crfs', *arXiv preprint arXiv:1412.7062* .

Cheshire, P. & Dericks, G. (2014), ''iconic design' as deadweight loss: rent acquisition by design in the constrained london office market'.

Chetty, R. & Hendren, N. (2018), 'The impacts of neighborhoods on intergenerational mobility ii: County-level estimates', *The Quarterly Journal of Economics* **133**(3), 1163–1228.

Chetty, R., Hendren, N. & Katz, L. F. (2016), 'The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment', *American Economic Review* **106**(4), 855–902.

Chuang, Y.-C., Ennett, S. T., Bauman, K. E. & Foshee, V. A. (2005), 'Neighborhood influences on adolescent cigarette and alcohol use: Mediating effects through parent and peer behaviors', *Journal of health and social behavior* **46**(2), 187–204.

Chung, H. L. & Steinberg, L. (2006), 'Relations between neighborhood factors, parenting behaviors, peer deviance, and delinquency among serious juvenile offenders.', *Developmental psychology* **42**(2), 319.

Chyn, E. (2018), 'Moved to opportunity: The long-run effects of public housing demolition on children', *American Economic Review* **108**(10), 3028–56.

Clendinning, J. P. (2002), 'Postmodern architecture/postmodern music', *Postmodern music/postmodern thought* pp. 119–40.

Coban, O., Ozyildirim, B. M. & Ozel, S. A. (2018), 'An empirical study of the extreme learning machine for twitter sentiment analysis', *International Journal of Intelligent Systems and Applications in Engineering* **6**(3), 178–184.

Coppersmith, G., Dredze, M., Harman, C. & Hollingshead, K. (2015), From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses, *in* 'Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality', pp. 1–10.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. & Mitchell, M. (2015), Clpsych 2015 shared task: Depression and ptsd on twitter, *in* 'Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality', pp. 31–39.

Coppersmith, G., Harman, C. & Dredze, M. (2014), Measuring post traumatic stress disorder in twitter, *in* 'Eighth International AAAI Conference on Weblogs and Social Media'.

Coppersmith, G., Ngo, K., Leary, R. & Wood, A. (2016), Exploratory analysis of social media prior to a suicide attempt, *in* 'Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology', pp. 106–117.

Cunha, F. & Heckman, J. (2007), 'The technology of skill formation', *American Economic Review* **97**(2), 31–47.

Curini, L., Iacus, S. & Canova, L. (2015), 'Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case', *Social Indicators Research* **121**(2), 525–542.

Cutler, D. M. & Glaeser, E. L. (1997), 'Are ghettos good or bad?', *The Quarterly Journal of Economics* **112**(3), 827–872.

Damm, A. P. & Dustmann, C. (2014), 'Does growing up in a high crime neighborhood affect youth criminal behavior?', *American Economic Review* **104**(6), 1806–32.

Daras, K. & Barr, B. (2020), 'Small area mental health index (samhi)'.

Del Bello, C. L., Patacchini, E. & Zenou, Y. (2015), 'Neighborhood effects in education'.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', Ieee, pp. 248–255.

Dishion, T. J. & McMahon, R. J. (1998), 'Parental monitoring and the prevention of child and adolescent problem behavior: A conceptual and empirical formulation', *Clinical Child and Family Psychology Review* **1**(1), 61–75.

Dubey, A., Naik, N., Parikh, D., Raskar, R. & Hidalgo, C. (2016), 'Deep learning the city : Quantifying urban perception at a global scale', *European Conference on Computer Vision (ECCV)* .

Dzogang, F., Lightman, S. & Cristianini, N. (2017), 'Circadian mood variations in twitter content', *Brain and Neuroscience Advances* **1**, 2398212817744501.

Edin, P.-A., Fredriksson, P. & Åslund, O. (2003), 'Ethnic enclaves and the economic success of immigrants—evidence from a natural experiment', *The Quarterly Journal of Economics* **118**(1), 329–357.

Elliott, M. C., Leventhal, T., Shuey, E. A., Lynch, A. D. & Coley, R. L. (2016), 'The home and the 'hood: Associations between housing and neighborhood contexts and adolescent functioning', *Journal of Research on Adolescence* **26**(1), 194–206.

Eriksen, M. D. (2010), 'Homeownership subsidies and the marriage decisions of low-income households', *Regional Science and Urban Economics* **40**(6), 490–497.

Folkman, S. (1984), 'Stress: appraisal and coping'.

Frölich, M. & Huber, M. (2014), 'Direct and indirect treatment effects: causal chains and mediation analysis with instrumental variables'.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. & Li, F. (2017), 'Using deep learning and google street view to estimate the demographic makeup of neighbourhoods across the united states', *PNAS* .

Gibbons, S., Peng, C., Tang, C. K. et al. (2019), Valuing the environmental benefits of canals using house prices, Technical report, Centre for Economic Performance, LSE.

Gibbons, S., Silva, O. & Weinhardt, F. (2013), 'Everybody needs good neighbours? evidence from students' outcomes in england', *The Economic Journal* **123**(571), 831–874.

Girshick, R. (2015), 'Fast r-cnn', *IEEE International Conference on Computer Vision* .

Glaeser, E. & Gyourko, J. (2018), 'The economic implications of housing supply', *Journal of Economic Perspectives* **32**(1), 3–30.

Gonçalves, A. P., Zuanazzi, A. C., Salvador, A. P., Jaloto, A., Pianowski, G. & Carvalho, L. d. F. (2020), 'Preliminary findings on the associations between mental health indicators and social isolation during the covid-19 pandemic', *Archives of Psychiatry and Psychotherapy* **22**(2), 10–19.

Google (2018), https://www.maps.google.com/.

Goux, D. & Maurin, E. (2007), 'Close neighbours matter: Neighbourhood effects on early performance at school', *The Economic Journal* **117**(523), 1193–1215.

Greenman, E., Bodovski, K. & Reed, K. (2011), 'Neighborhood characteristics, parental practices and children's math achievement in elementary school', *Social science research* **40**(5), 1434–1444.

Hamza, C. A., Ewing, L., Heath, N. L. & Goldstein, A. L. (2021), 'When social isolation is nothing new: A longitudinal study on psychological distress during covid-19 among university students with and without preexisting mental health concerns.', *Canadian Psychology/Psychologie canadienne* **62**(1), 20.

Han, J. (2021), Parental involvement and neighborhood quality: Evidence from public housing demolitions in chicago, Working paper.

Hartig, T., Evans, G. W., Jamner, L. D., Davis, D. S. & Gärling, T. (2003), 'Tracking restoration in natural and urban field settings', *Journal of Environmental Psychology* **23**(2), 109–123.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the Econometric Society* pp. 153–161.

Heckman, J. J. & Pinto, R. (2015), 'Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs', *Econometric Reviews* **34**(1-2), 6–31.

Hilber, C. A. & Vermeulen, W. (2016), 'The impact of supply constraints on house prices in england', *The Economic Journal* **126**(591), 358–405.

Hillier, A. E. (2003), 'Redlining and the home owners' loan corporation', *Journal of Urban History* **29**(4), 394–420.

Hillier, A. E. (2005), 'Residential security maps and neighborhood appraisals: The home owners' loan corporation and the case of philadelphia', *Social Science History* pp. 207–233.

Homan, C., Johar, R., Liu, T., Lytle, M., Silenzio, V. & Alm, C. O. (2014), Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale, *in* 'Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality', pp. 107–117.

Hsieh, C.-T. & Moretti, E. (2019), 'Housing constraints and spatial misallocation', *American Economic Journal: Macroeconomics* **11**(2), 1–39.

Hu, M. & Wang, X. (2019), 'Homeownership and household formation: no homeownership, no marriage?', *Journal of Housing and the Built Environment* pp. 1–19.

Huang, B. & Carley, K. M. (2019), A large-scale empirical study of geotagging behavior on twitter, *in* 'Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining', pp. 365–373.

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X. & Zhu, T. (2014), Detecting suicidal ideation in chinese microblogs with psychological lexicons, *in* '2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops', IEEE, pp. 844–849.

Janssens, O., Verstockt, S., Mannens, E., Van Hoecke, S. & Van de Walle, R. (2014), Influence of weak labels for emotion recognition of tweets, *in* 'Mining Intelligence and Knowledge Exploration', Springer, pp. 108–118.

Johnson, R. C. & Jackson, C. K. (2019), 'Reducing inequality through dynamic complementarity: Evidence from head start and public school spending', *American Economic Journal: Economic Policy* **11**(4), 310–49.

Kahneman, D. & Krueger, A. B. (2006), 'Developments in the measurement of subjective well-being', *Journal of Economic Perspectives* **20**(1), 3–24.

Kalamatianos, G., Symeonidis, S., Mallis, D. & Arampatzis, A. (2018), 'Towards the creation of an emotion lexicon for microblogging', *Journal of Systems and Information Technology* **20**(2), 130–151.

Katz, L. F., Kling, J. R. & Liebman, J. B. (2001), 'Moving to opportunity in boston: Early results of a randomized mobility experiment', *The Quarterly Journal of Economics* **116**(2), 607–654.

Keele, L., Titiunik, R. & Zubizarreta, J. R. (2015), 'Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 223–239.

Killgore, W. D., Cloonan, S. A., Taylor, E. C. & Dailey, N. S. (2020), 'Loneliness: A signature mental health concern in the era of covid-19', *Psychiatry Research* **290**, 113117.

Kim, H. H.-s. & Jung, J. H. (2021), 'Social isolation and psychological distress during the covid-19 pandemic: A cross-national analysis', *The Gerontologist* **61**(1), 103–113.

Kim, S., Bak, J. & Oh, A. H. (2012), Do you feel what i feel? social aspects of emotions in twitter conversations, *in* 'Sixth International AAAI Conference on Weblogs and Social Media'.

Kling, J. R., Liebman, J. B. & Katz, L. F. (2007), 'Experimental analysis of neighborhood effects', *Econometrica* **75**(1), 83–119.

Kling, J. R., Liebman, J. B., Katz, L. F. & Sanbonmatsu, L. (2004), 'Moving to opportunity and tranquility: Neighborhood effects on adult economic self-sufficiency and health from a randomized housing voucher experiment', *Available at SSRN 588942* .

Kloosterman, R. C. & Musterd, S. (2001), 'The polycentric urban region: towards a research agenda', *Urban studies* **38**(4), 623–633.

Kondo, A. & Shoji, M. (2019), 'Peer effects in employment status: Evidence from housing lotteries', *Journal of Urban Economics* **113**, 103195.

Krimmel, J. (2018), 'Persistence of prejudice: Estimating the long term effects of redlining'.

Krizhevsky, A., Sutskever, I. & Hinton, G. (2012), 'Imagenet classification with deep convolutional neural networks', *Advances in neural information processing* .

Law, S., Seresinhe, C. I., Shen, Y. & Gutierrez-Roig, M. (2018), 'Street-frontage-net: urban image classification using deep convolutional neural networks', *International Journal of Geographical Information Science* **0**(0), 1–27.
URL: *https://doi.org/10.1080/13658816.2018.1555832*

Lepore, S. J. & Evans, G. W. (1996), 'Coping with multiple stressors in the environment.'.

Leventhal, T. & Brooks-Gunn, J. (2003), 'Moving to opportunity: an experimental study of neighborhood effects on mental health', *American Journal of Public Health* **93**(9), 1576–1582.

Leventhal, T. & Brooks-Gunn, J. (2005), 'Neighborhood and gender effects on family processes: Results from the moving to opportunity program', *Family Relations* **54**(5), 633–643.

Lofland, L. H. (2017), *The public realm: Exploring the city's quintessential social territory*, Routledge.

Loos, A. (2019), *Ornament and crime*, Penguin UK.

Loveys, K., Crutchley, P., Wyatt, E. & Coppersmith, G. (2017), Small but mighty: affective micropatterns for quantifying mental health from social media language, *in* 'Proceedings of the

Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality', pp. 85–95.

Lucchetti, G., Góes, L. G., Amaral, S. G., Ganadjian, G. T., Andrade, I., de Araújo Almeida, P. O., do Carmo, V. M. & Manso, M. E. G. (2020), 'Spirituality, religiosity and the mental health consequences of social isolation during covid-19 pandemic', *The International Journal of Social Psychiatry* .

Ma, J., Hua, T., Zeng, K., Zhong, B., Wang, G. & Liu, X. (2020), 'Influence of social isolation caused by coronavirus disease 2019 (covid-19) on the psychological characteristics of hospitalized schizophrenia patients: a case-control study', *Translational Psychiatry* **10**(1), 1–5.

MacKerron, G. & Mourato, S. (2020), Mappiness: natural environments and in-the-moment happiness, *in* 'Handbook on Wellbeing, Happiness and the Environment', Edward Elgar Publishing.

Malamud, O., Pop-Eleches, C. & Urquiola, M. (2016), Interactions between family and school environments: Evidence on dynamic complementarities?, Technical report, National Bureau of Economic Research.

Manson, S., Schroeder, J., Van Riper, D. & Ruggles, S. (2019), 'Ipums national historicla geographical information system: Version 14.0 [database]. ipums', *Institute for Social Research and Data Innovation. University of Minnesota* .

Massey, D. S. (2015), The legacy of the 1968 fair housing act, *in* 'Sociological Forum', Vol. 30, Wiley Online Library, pp. 571–588.

Mautong, H., Gallardo-Rumbea, J. A., Alvarado-Villa, G. E., Fernández-Cadena, J. C., Andrade-Molina, D., Orellana-Román, C. E. & Cherrez-Ojeda, I. (2021), 'Assessment of depression, anxiety and stress levels in the ecuadorian general population during social isolation due to the covid-19 outbreak: a cross-sectional study', *BMC Psychiatry* **21**(1), 1–15.

McDonell, J. R. (2007), 'Neighborhood characteristics, parenting, and children's safety', *Social Indicators Research* **83**(1), 177–199.

McGonagle, K. A. & Kessler, R. C. (1990), 'Chronic stress, acute stress, and depressive symptoms', *American Journal of Community Psychology* **18**(5), 681–706.

McManus, K., Mallory, E. K., Goldfeder, R. L., Haynes, W. A. & Tatum, J. D. (2015), 'Mining twitter data to improve detection of schizophrenia', *AMIA Summits on Translational Science Proceedings* **2015**, 122.

Miller, J. J. & Park, K. A. (2018), 'Same-sex marriage laws and demand for mortgage credit', *Review of Economics of the Household* **16**(2), 229–254.

Mitchell, M., Hollingshead, K. & Coppersmith, G. (2015), Quantifying the language of schizophrenia in social media, *in* 'Proceedings of the 2nd workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality', pp. 11–20.

Mohammad, S. M. (2012), # emotional tweets, *in* 'Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation', Association for Computational Linguistics, pp. 246–255.

Mohammad, S. M. & Bravo-Marquez, F. (2017), 'Emotion intensities in tweets', *arXiv preprint arXiv:1708.03696* .

Mohammad, S. M. & Yang, T. W. (2011), Tracking sentiment in mail: How genders differ on emotional axes, *in* 'Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis', Association for Computational Linguistics, pp. 70–79.

Murayama, H., Okubo, R. & Tabuchi, T. (2021), 'Increase in social isolation during the covid-19 pandemic and its association with mental health: Findings from the jacsis 2020 study', *International Journal of Environmental Research and Public Health* **18**(16), 8238.

Myrtek, M. & Spital, S. (1986), 'Psychophysiological response patterns to single, double, and triple stressors', *Psychophysiology* **23**(6), 663–671.

Naik, N., Philipoom, J., Raskar, R. & Hidalgo, C. (2014), Streetscore - predicting the perceived safety of one million streetscapes, *in* 'CVPR Workshop on Web-scale Vision and Social Media'.

Oreopoulos, P. (2003), 'The long-run consequences of living in a poor neighborhood', *The Quarterly Journal of Economics* **118**(4), 1533–1575.

Paciorek, A. (2013), 'Supply constraints and housing market dynamics', *Journal of Urban Economics* **77**, 11–26.

Pak, A. & Paroubek, P. (2010), Twitter as a corpus for sentiment analysis and opinion mining., *in* 'LREc', Vol. 10, pp. 1320–1326.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J. & Brew, C. (2012), 'Sentiment analysis of suicide notes: A shared task', *Biomedical Informatics Insights* **5**, BII–S9042.

Rauschenberg, C., Schick, A., Goetzl, C., Roehr, S., Riedel-Heller, S. G., Koppe, G., Durstewitz, D., Krumm, S. & Reininghaus, U. (2021), 'Social isolation, mental health, and use of digital interventions in youth during the covid-19 pandemic: A nationally representative survey', *European Psychiatry* **64**(1).

Resch, B., Summa, A., Sagl, G., Zeile, P. & Exner, J.-P. (2015), Urban emotions—geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data, *in* 'Progress in location-based services 2014', Springer, pp. 199–212.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A. & Boyd-Graber, J. (2015), Beyond lda: exploring supervised topic modeling for depression-related language in twitter, *in* 'Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality', pp. 99–107.

Ricks, J. S. (2021), 'Mortgage subsidies, homeownership, and marriage: Effects of the va loan program', *Regional Science and Urban Economics* **87**, 103650.

Robb, C. E., de Jager, C. A., Ahmadi-Abhari, S., Giannakopoulou, P., Udeh-Momoh, C., McKeand, J., Price, G., Car, J., Majeed, A., Ward, H. et al. (2020), 'Associations of social isolation with anxiety and depression during the early covid-19 pandemic: a survey of older adults in london, uk', *Frontiers in Psychiatry* **11**.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J. & Harabagiu, S. M. (2012), Empatweet: Annotating and detecting emotions on twitter., *in* 'Lrec', Vol. 12, Citeseer, pp. 3806–3813.

Rotger, G. P. & Galster, G. C. (2019), 'Neighborhood peer effects on youth crime: natural experimental evidence', *Journal of Economic Geography* **19**(3), 655–676.

Rowe, H. A. (2011), 'The rise and fall of modernist architecture', *Inquiries Journal* **3**(04).

Saha, K., Chan, L., De Barbaro, K., Abowd, G. D. & De Choudhury, M. (2017), 'Inferring mood instability on social media by leveraging ecological momentary assessments', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**(3), 95.

Sanbonmatsu, L., Katz, L. F., Ludwig, J., Gennetian, L. A., Duncan, G. J., Kessler, R. C., Adam, E. K., McDade, T. & Lindau, S. T. (2011), 'Moving to opportunity for fair housing demonstration program: Final impacts evaluation'.

Seresinhe, C., Preis, T. & Moat, S. (2017), 'Using deep learning to quantify the beauty of oudoor places', *Royal Society Open Science* .

Simonyan, K. & Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition', *arXiv:1409.1556* .

Sintsova, V., Musat, C. & Pu, P. (2014), Semi-supervised method for multi-category emotion recognition in tweets, *in* '2014 IEEE International Conference on Data Mining Workshop', IEEE, pp. 393–402.

Smith, B. M., Twohy, A. J. & Smith, G. S. (2020), 'Psychological inflexibility and intolerance of uncertainty moderate the relationship between social isolation and mental health outcomes during covid-19', *Journal of Contextual Behavioral Science* **18**, 162–174.

Strapparava, C. & Mihalcea, R. (2008), Learning to identify emotions in text, *in* 'Proceedings of the 2008 ACM Symposium on Applied Computing', ACM, pp. 1556–1560.

Suttles, J. & Ide, N. (2013), Distant supervision for emotion classification with discrete binary values, *in* 'International Conference on Intelligent Text Processing and Computational Linguistics', Springer, pp. 121–136.

Tendulkar, S. A., Buka, S., Dunn, E. C., Subramanian, S. & Koenen, K. C. (2010), 'A multilevel investigation of neighborhood effects on parental warmth', *Journal of Community Psychology* **38**(5), 557–573.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica: Journal of the Econometric Society* pp. 24–36.

VOA (2010), https://www.gov.uk/guidance/how-non-domestic-property-including-plant-and-machinery-is-valued.

Watson, D. & Clark, L. A. (1997), 'Measurement and mismeasurement of mood: Recurrent and emergent issues', *Journal of Personality Assessment* **68**(2), 267–296.

Weinhardt, F. (2014), 'Social housing, neighborhood quality and student performance', *Journal of Urban Economics* **82**, 12–31.

Zuberi, A. (2016), 'Neighborhoods and parenting: Assessing the influence of neighborhood quality on the parental monitoring of youth', *Youth & Society* **48**(5), 599–627.