Approaches to Estimating the Value of Priority Admission to Primary Schools in Singapore

Benjamin Chow

Quantitative Methods in the Social Sciences
Graduate School of Arts and Sciences
Columbia University

Advisor: Jeremy Porter, QMSS, Columbia University

May 2015

I am grateful to my adviser, Jeremy Porter, for his comments, feedback and encouragement throughout this endeavor. I would also like to thank Stephen Machin, Sung-Woo Cho, Randall Rebeck, Elizabeth Kopko, Claire Kelley, Benjamin Singer and Kean Yung Kwek for substantive comments which helped to shape my work;Yam Yujian, Benjamin Ng, Helen Khoo, and Ng Mui Leng from the Urban Redevelopment Authority of Singapore for their support during this process. Any errors are my own.

Abstract

Regression discontinuity has become an increasingly popular tool for estimating the value of good schools on housing prices. In Singapore, families living within arbitrarily-defined 1km and 2km threshold distances around primary schools in Singapore enjoy priority privileges in registration exercises for incoming Primary One students,[1] which provides a seemingly ideal set-up for a discontinuity design. However, results from both parametric and nonparametric discontinuity designs suffer from inconsistency across model specification, and diagnostic checks reveal spatial trends to be interfering with the results. A spatial matching approach finds significant premiums of 3.6% and 1.0% for private housing at the 1km threshold and public housing at the 2km threshold, around schools featuring the Gifted Education Programme.

[1] "Primary schools" are the equivalent of "elementary schools" in other parts of the world. A full list of the terminology used in this paper, which is in accordance with the schooling system in Singapore, is provided in Appendix A.

1. Introduction

Every year in Singapore, children turning age 7 in the coming year,[2] along with their parents, participate in a registration exercise for admission to primary schools. The exercise allocates children to various primary schools across Singapore, with the aim of achieving a balance in student composition through according priority to certain children based on certain criteria. These include those with siblings or parents from the same school (preservation of school culture), those with parents serving as school volunteers or as staff members (rewards for service), those not in any of the above categories (diversity), and those who live near to the school (practical reasons such as higher student involvement).[3] This paper focuses on quantifying the value of priority on the basis of the last reason, since by living in properties close to schools of their choice, parents thus improve the chances of being able to enrol their children into these schools.

A select handful of the 187 primary schools in Singapore are typically more sought-after during the registration exercise. These are those that consistently outperform the national average in terms of test scores achieved by their students in the Primary School Leaving Examination (PSLE), a standardized exam for students aged 12 in Primary Six. As high school placement is largely determined by this examination, enrolment in popular schools is competitive,[4] and where applications exceed vacancies a ballot will be conducted. The distance-based priority affects enrolment probability within each phase,[5] as priority is first given to those staying within a 1km (0.62 mi) of the school, followed by those staying between the 1km and 2km boundaries. Consistent with the literature on this topic,[6] I hypothesize that a fraction of the value of properties located within boundaries of better

[2] The school year in Singapore begins every January.
[3] In the National Day Rally 2013, Prime Minister of Singapore Lee Hsien Loong highlighted that recent changes to increase the fixed number of places were with the intent of "striking a balance."
[4] See recent news articles in local newspaper the Straits Times by Chia (2013) and Lee (2014).
[5] A table detailing the different registration phases and priority levels is provided in Appendix B1.
[6] Black and Machin (2010) count 54 studies using this methodology as of their time of publication.

schools is attributable to what I call the priority admission value (PAV),[7] while the PAV should be 0 for undersubscribed schools. Against this backdrop, this paper is concerned with the following research questions:

- Are homebuyers willing to pay a premium to gain priority admission to good schools?
- Can a regression discontinuity identify the value of the PAV? If so, how should space be handled as a variable that influences housing prices?

As reflected in the ensuing literature review, some studies argue that PAV can be indirectly measured through exploiting variation in eligibility rules across administrative boundaries.[8] Others use housing prices as a channel to measure parental willingness-to-pay for school quality, as measured through other proxies.[9] Even if a Regression Discontinuity Design (Sections 2.3 and 3.4) is able to isolate the PAV from spatial characteristics, the PAV is by definition affected by both school quality and the probability of admission. Some external studies referred to by this paper assume this probability to be a constant, at one; however, in Singapore, high housing density means that even living within a 1km boundary does not guarantee admission.[10] Thus, the PAV should be interpreted as a combination of school quality value as well as factors influencing probability of admission.

The second research question asks if geographic space, as a variable, can really be controlled for effectively using the discontinuity design. Many of the studies surveyed in the literature using this methodology employ crude controls for spatial dependence, such as clustered standard errors and town or district fixed effects. In the ideal discontinuity design,

[7] The value is prospective because parents may buy properties in anticipation of requiring the proximity for their children years down the road, yet it is impossible to know the over-subscription rate until the exercise is over, as well as which of the admission priority criteria will prove crucial when the time comes.
[8] This includes one of the two first studies utilizing the discontinuity approach, by Bogart & Cromwell (1997).
[9] The most commonly used proxies are measures of test scores, such as in Black (1999) and Fack & Grenet (2010), which may feature either in absolute levels or in terms of value-added. A short discussion on measures of school quality is provided in the following section.
[10] As revealed in the recent 2014 Primary One Registration Exercise, even proximity to school alone (Phase 2C) is not enough to guarantee a spot for a child (Chia, 2013; Lee, 2014), as other forms of priority aforementioned factor into the equation as well.

spatial variables would theoretically be differenced away by selecting a fine enough bandwidth around the boundary, so their inclusion should leave estimates invariant. However, distances between properties can never be zero, and local spatial factors such as amenities can influence housing prices greatly. The final research question of this study is concerned with exploring this hypothesis through the use of different estimation strategies as outlined in the spatial discontinuity literature.

The first research question has important implications for education policy. The Ministry of Education's initiative to make "Every School a Good School" (MOE, 2013) was largely a response to parental obsession with PSLE scores and getting their children into good primary schools (see PMO, 2013). Identifying and understanding factors that influence parental demand for schooling independent of neighborhood characteristics will be key in managing oversubscription in popular schools.

On the theoretical side, this paper makes contributions to two existing domains of academia. The first is the debate on the appropriateness of different measures of school quality. Given that education results in such multi-dimensional outcomes, it can be argued that parental valuation is a strong indicator of overall school quality, since as consumers of (primary) education they tend to take into account more factors than are observable through data. A revealed preference argument would be that knowledge of their valuation of schools enables us to determine what they consider a high-quality education to be. The second is the expanding body of literature on regression discontinuity approaches. Exploiting discontinuity in a geographic setting invites an additional source of endogeneity from spatial factors. Many studies exploring the relationship between housing prices and education quality reviewed in the course of this paper did not attempt to overtly model or control for spatial dependence.[11]

[11] Notable exceptions to this are Gibbons & Machin (2003), Gibbons, Machin & Silva (2013) and Keele & Titiunik (2015).

Through a series of robustness checks I seek to determine the possible pitfalls from aspatial approaches and to evaluate the strength of spatial discontinuity approaches.

The outline of this paper is as follows: Section 2 provides a review of the literature on measuring school quality through housing prices; Section 3 introduces the datasets, methodology and the econometric models employed; Section 4 presents descriptive statistics and preliminary visualizations of the data; Section 5 presents the actual results from the models, which are discussed in Section 6. Section 7 concludes.

2. Literature Review

2.1 Measuring School Quality

What makes a good school? As of February 2015, a cursory Google™ search produces 1.2 billion websites claiming to answer this question; however if there is one thing the education literature agrees on, it is the elusiveness of a quantifiable measure of education quality. The most oft-used definitions and measures of school quality – dichotomized into inputs[12] and outputs[13] – each suffer from an array of measurement error and omitted variables bias, and do not even always correlate strongly with one another. Meanwhile, attempts at more precise measurement such as classroom observation[14] run into cost-related problems and teacher opposition for implementation at a larger scale.

Resultantly, many researchers have instead focused on using alternative quasi-experimental approaches in studies of education, whether in quantifying the impact of specific

[12] The most common examples are per-pupil spending and teacher quality. For spending, see Clotfelter, Ladd, & Vigdor (2010) and Taylor & Fowler (2006) for specific examples of issues faced, while Ladd & Loeb (2014) provide a thorough evaluation. Teacher quality is measured either through a context-specific scoring metric (Ferguson & Brown, 2000; Araujo et al., 2014), relative wages (Loeb & Page, 2000), selectivity of the college the teachers attended (Ehrenberg & Brewer, 1994), or, most frequently, attributed to elusive unobservable factors (Hanushek, 1986; Rivkin, Kain & Hanushek, 2005; Aaronson, Barrow & Sander, 2007; Hanushek, 2005).
[13] The most common short-run outcomes are test scores and drop-out rates, while a battery of long-run outcomes have been analyzed, including the probability of imprisonment (Lochner & Moretti, 2004), health (Lochner & Moretti, 2004; Lleras-Muney, 2005, Meghir, Palme & Simeonova, 2012), voter turnout (Milligan, Moretti, & Oreopoulos, 2003), children's earnings and test scores (Black, Devereux, & Salvanes, 2005; Carneiro, Meghir, & Parey, 2013), and social returns (Acemoglu & Angrist, 2000; Ciccone & Peri, 2002; Moretti, 2004)
[14] A good example of this methodology is illustrated in Kane, Taylor, Tyler, & Wooten (2011).

inputs,[15] the returns to education,[16] or measuring education quality itself. One such methodology exploits the capitalization of school value into property prices, which has traditionally been measured using hedonic pricing models (Rosen, 1974; Griliches, 1990) and more recently, by using a regression discontinuity[17] approach.

2.2 Problems in Measuring School Quality through Housing Prices

In seeking to infer the value of schools from variation in residential property prices, researchers typically encounter problems of endogeneity. These arise through the omission of unobserved variables from regression analyses, and prevent identification of school value.

The biggest source of endogeneity is neighborhood sorting (Tiebout, 1956). Good schools are typically located in neighborhoods with more desirable attributes, such as prestige, security, cleanliness and better local amenities. As such, parents buying into districts where good schools are located pay for the value of these other unobserved attributes as well, making it difficult to disentangle school value from the value of the neighborhood and its associated public goods and services. Sheppard's (1999) survey notes that many studies do not account for land value as a function of location, and some even omit both variables.

Gibbons & Machin (2003) raise an additional source of endogeneity that occurs in the US, where the tax base of the district determines school funding. Thus, schools in wealthier neighborhoods receive more funding. Although this does not automatically translate into better schooling outcomes, it is still a nontrivial issue explored by researchers (see for instance, Bogart & Cromwell, 1997). This problem does not arise in Singapore, where the Ministry of Education (MOE) determines funding centrally, so we focus our attention on the former issue of sorting.

[15] See Card & Krueger (1996) and Angrist & Lavy (1999)
[16] See Angrist & Krueger (1991), Ashenfelter & Krueger (1994) and Carneiro, Heckman, & Vytlacil (2011).
[17] The origins of this methodology are traced back to its first occurrence in Thistlewaite & Campbell (1960), termed "Regression-Discontinuity Analysis."

A final issue in analyses of housing prices is that of spatial dependence, which refers to the lack of independence across observations in spatial data (Cliff & Ord, 1973; Anselin, 2001). Housing is intrinsically spatial in nature, and as will be demonstrated below with my own dataset, property prices are typically very significantly correlated across space. The dependence filters through in terms of neighborhood value, but also as a result of localized geographic characteristics, such as proximity to public transportation, and development characteristics (i.e., due to the fact that each construction project usually comprises several apartment blocks in close proximity).

2.3 Regression Discontinuity Designs (RDDs)

In this section I show how the RDD is able to overcome the potential problems as outlined above. Its premise is that primary school admissions are often contingent on pre-defined attendance boundaries, such that households on either side of the boundary face differing probabilities of entry to specific schools. Such housing units are typically in the same neighborhood, and will have approximately the same locational value, being virtually next to each other. Thus, a comparison of housing prices on either side of the boundary must reflect an implicit estimation of school quality valuation as long as all other attributes can be assumed to be approximately constant (Black, 1997).

Can all other attributes be assumed to be approximately constant? Apart from spatial attributes, there are other factors influencing the value of properties at a local level. Physical characteristics of individual housing units, such as floor space, number of rooms, or floor level, also influence housing value, and provide the impetus for the theoretical precursor to the RDD: hedonic pricing models (Rosen, 1974; Griliches, 1990). On its own, a hedonic pricing model is based on the premise that properties are bundles of goods whose value can be decomposed into deriving from observed dwelling characteristics. The main drawback of

the model is its inability to control for spatial attributes.[18] In the RDD, physical characteristics are easily controlled for, effectively embedding a hedonic pricing approach within.

The earliest most well-known application of an RDD in a geographic context is by Black (1999), with two other works by Brasington & Haurin (1996) and Bogart & Cromwell (1997) employing similar estimation methodologies, but without full justification as discontinuity designs. Since then, many studies have used RDDs, and others have provided guidance on how to evaluate and justify the approach.[19]

2.4 RDDs and Spatial effects

The domain of spatial econometrics has evolved in recent decades, providing guidance on the treatment of geographic space as an omitted variable. In this context, the aim of incorporating the techniques introduced by the spatial statistical literature is solely to control for space. As such, the construction of a full-fledged geo-statistical model is beyond the scope of this paper. For a general treatment of spatial dependence and the techniques involved, I refer the reader to Anselin's works on spatial econometrics (Anselin, 1988, 2001, 2009).

Many studies utilizing RDDs in a geographic setting, including Black (1999), claim that the discontinuity approach eliminates spatial variables if the bandwidth chosen is small enough, and that robust or clustered standard errors (Wong, 2011; Gibbons, McNally, & Viarengo, 2013) are sufficient to address this problem. However, recent papers by Gibbons, Machin, & Silva (2013) and Keele & Titiunik (2015) argue that more should be done to justify the discontinuity approach, and provide guidance on ways to do this.

[18] Sheppard's (1999) review of the literature on hedonic models reveals "very few [hedonic pricing] models explicitly incorporate a land value function that depends upon location." Gibbons & Machin (2003) bring up the possibility of using postcode districts as a crude control for location, but reject it on the basis that there exists "no theoretical basis for believing that these are the right controls," and that imperfect controls in general will lead to inconsistent estimates.

[19] See Bayer, Ferreira, & McMillan (2007), Gibbons & Machin (2008), Fack & Grenet (2010), Machin (2011), Wong (2011), and Gibbons, Machin, & Silva (2013) for examples.

Anselin (2001) identifies three main approaches to control for space, but singles out the specification of a spatial stochastic process as the most viable. The most straightforward approach is to specify a spatial lag model, where the lag is defined as a weighted average of the dependent variable of observations in a given neighborhood. The key choices here are the magnitude of the spatial *lag,* which is the neighborhood of observations, and the weighting scheme used. In the only study surveyed which specified some sort of spatial model explicitly, Gibbons and Machin (2003) experimented with different choices of bandwidth, remarking that there is "no way of knowing, other than casual empiricism, what geographical area comprises the correct reference group" (p. 203). As for the weighting scheme, a common choice in the spatial econometrics literature is to use Inverse-Distance Weighted lags (for instance, see Shepard, 1968).

2.5 Studies in Singapore

In Singapore, two studies have aimed to measure school quality through housing prices. The first, by Agarwal, Rengarajan & Sing (2014), uses difference-in-difference estimation exploiting the incidence of school relocation events. Another study, more closely related to mine, is Wong (2011), which employs a similar strategy in estimating the PAV.

This study uses Wong's (2011) as a reference point, but departs from it along several lines. Firstly, his study uses public housing (HDB flats) data, while I use an additional dataset on private property transactions in addition to HDB price data. It will be interesting to compare if there are differences in PAVs between those living in HDB flats and private property. Secondly, Wong (2011) employs a traditional hedonic pricing estimation method, restricting the estimation to various bandwidths around the school boundaries. He controls for HDB town fixed effects to account for omitted spatial variables, which has an effect of lowering his parameter estimates. In this paper, I include spatially weighted prices as a covariate to control for space, in addition to experimenting with different bandwidths.

3. Research Methodology

3.1 Data[20]

This study involves several categories of data. The first is housing transactions data, which are divided up into data on public and private housing.[21] Private housing transaction data are taken from REALIS, an online portal maintained by the Urban Redevelopment Authority of Singapore. Data on public housing transaction data was made publicly available as of December 2014. A final category is spatial data, which includes geo-coded references and shapefiles of Singapore with its planning boundaries, and is also made publicly available.

Housing transactions were joined to geo-references and mapped onto the shapefile of Singapore using QGIS software. An important institutional characteristic of the distance-based priority criterion is that the boundaries are exactly circular and set at arbitrary distances of 1km and 2km from each school. This simplifies the analysis somewhat, since proximity to the boundary and proximity to the school are captured by a single unambiguous distance measure, precluding the need for examination of the interaction between school distance and boundary distance.[22]

The next issue is to define a subset of schools to run the analyses on. The MOE prefers not to make most official data on schools, such as those on test scores and school inputs, publicly available. The only data they publish are oversubscription rates for the most recent registration exercise, on-line (The Straits Times, 2014). As such, I obtained various rankings of schools from educational consultancies and watchdogs that compile and publish unofficial statistics for prospective parents.[23]

[20] Data referred to as "publicly available" can be referenced from http://data.gov.sg, a portal which houses links to publicly-available datasets from various Singapore government ministries and agencies.
[21] See Appendix B2 for clarifications on the definitions of "public" and "private" housing.
[22] Geo-coded references utilized a WGS84 projection, and QGIS was used to compute distances between housing units and the boundaries. Using Euclidean distance was found to be inaccurate due to the projection.
[23] Typically, one might be wary about the reliability of data from such sources. However, I make the following case why such data may still be useful: firstly, given the lack of publicly available data, parents searching for school rankings do have access and may use these rankings (regardless of their reliability) in making decisions

The set of schools I have chosen comprises primary schools running the Gifted Education Programme (GEP), henceforth referred to as "GEP Schools."[24] These schools have traditionally performed well in the PSLE, and the average school ranking shown in Table 3.1, as measured by top scoring students in each school, illustrates this showing even in recent years. I also compute a crude measure of schools' reputation scores using the aforementioned rankings, as a rough proxy for the *prestige* traditionally associated with these schools. More details about the rankings and methodology can be found in Appendices B3 and B4.

Table 3.1. Top 20 Primary Schools*

School	Gifted?	Rank (2014)	Rank (2013)	Rank (2012)	Avg. Rank	Reputation Score	Reputation Rank
Raffles Girls' Primary School	Yes	8	1	8	5.67	99.5	2
St. Hilda's Primary School	Yes	2	2	17	7	95	3
Nan Hua Primary School	Yes	13	4	6	7.67	60	10
Kuo Chuan Presbyterian Primary School	No	NA	9	8	8.5	2.5	38
Anglo-Chinese School (Primary)	Yes	1	17	10	9.33	81.5	6
Rosyth School	Yes	5	9	14	9.33	89.5	4
Nanyang Primary School	Yes	25	4	1	10	112.5	1
Tao Nan School	Yes	25	2	3	10	73	8
Ai Tong School	No	16	9	10	11.67	40	12
Catholic High School	Yes	34	4	5	14.33	78.5	7
South View Primary School	No	25	4	14	14.33	0	NA
St. Anthony's Primary School	No	16	NA	17	16.5	0	NA
Nan Chiau Primary School	No	16	17	20	17.67	26.5	16
Temasek Primary School	No	8	39	6	17.67	18	22
Rulang Primary School	No	13	39	3	18.33	64	9
Ngee Ann Primary School	No	16	24	NA	20	0	NA
Henry Park Primary School	Yes	13	39	17	23	83	5
Princess Elizabeth Primary School	No	34	17	NA	25.5	0	NA
Northland Primary School	No	NA	9	44	26.5	0	NA
Anglo-Chinese School (Junior)	No	NA	34	24	29	3	37

*** Lower ranks and higher scores represent better standing**

Notes on rank computation:
- Schools are ranked by scores of the highest scoring pupil, as reported on kiasuparents.com (Updated April 29, 201
- Standard competition ("1224") ranking is utilized, so tied observations receive the lowest rank
- Only schools with at least 2 rankings out of the 3 years are considered

Notes on reputation scores:
- School rankings from 7 unofficial sources were tabulated
- Each time a school featured in a ranking table it received a positive score, with a maximum of 20 for each ranking
- Full details about the score computation are provided in Appendix B

on where to buy housing; secondly, if such indicators were not useful they would arguably simply show up as insignificant in subsequent regressions; thirdly, oversubscription rates in these websites were cross-checked with statistics released in the local press as far as possible and found to be accurate. For instance, the Straits Times publishes an article after every phase of the registration process, typically reporting on schools with outlier over- or under- subscription rates and recent trends. The two articles aforementioned, Chia (2013) and Lee (2014), are a case in point. For a discussion on the use of these sources, see Heng (2013).

[24] See MOE (2014) for more information on the Gifted Education Programme.

3.2 Observations used

The estimation strategy (Section 3.4) relies on the comparability of housing units within and outside the boundaries to identify the PAV. Table 1 in Appendix A shows the specific numbers of observations dropped at each stage. Due to data availability for public housing transactions, I selected only private housing transactions from March 2012 to November 2014. In terms of housing types, only condominiums and apartment type housing (private housing) and housing with more than three rooms (public housing) were included in the dataset, as these tend to be more uniform in terms of physical attributes and comparable.[25] In comparison, floor area is often ambiguous for landed or terraced housing, since gardens, ditches or derelict parts of the house may be included in the floor area measurement. I also drop a small amount of observations with missing data on age from the private housing transactions dataset (3.3%), and outlier observations with floor area more than 300 square meters (1.3%, 3229 square feet). Details on the last point are provided in Appendix B.3.

3.3 Simplifications due to the Institutional Context in Singapore

From the literature, the most pressing concern in studies of housing prices and school quality is omitted variables bias arising from two sources: spatial effects and cross-boundary effects. Spatial effects are factors tied to geographic location that affect housing price, which include neighborhood attributes and other proximity measures such as to the city center or public transportation. Cross-boundary effects are characteristics other than school accessibility that change across the boundaries, and include boundary coincidence and neighborhood sorting effects. Four papers by Imbens and Lemieux (2008), Lee and Lemieux

[25] Based on the most recent 2010 census, 93% of households live in apartment-type housing (inclusive of both public and private housing), while excluding 1- and 2-room HDB flats brings the number down to 89%. As such, using this subset of housing data maintains both comparability (between private and public housing) and generalizability of results.

(2010), Gibbons, Machin, and Silva (2013) and Keele and Titiunik (2015)[26] provide guidelines to reinforce the credibility of the RDD. Institutional characteristics of education in Singapore allow for some simplifications to be made with regard to the RDD, so that not all the safeguards mentioned in the papers apply.

Keele & Titiunik (2015) raise the possibility that treatment effects[27] may be heterogeneous along the boundary. In this case, priority admission does not vary around the boundary but only across it. Moreover, the boundaries in question are circles at specific intervals of 1km and 2km, unlike school district boundaries in other countries that vary in shape. Thus, the forcing variable – distance from school – can be regarded uni-dimensionally.

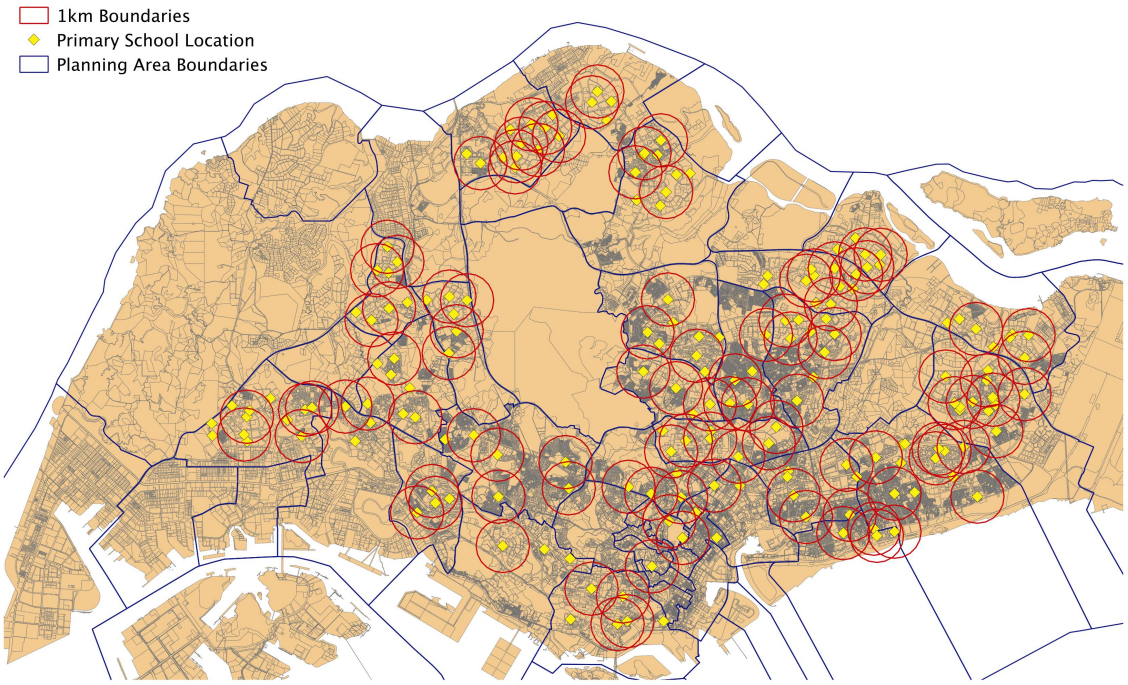
These circular boundaries present another simplification. The possibility of boundary coincidence is unlikely since the circles drawn are exogenous, defined according to geometry and not according to regional characteristics. As such, there is little likelihood of compound treatment effects[28] highlighted by Keele & Titiunik (2015) and Gibbons, Machin & Silva (2013). An illustration is provided in Figure 3.2 (inset), showing a map of Singapore and its 55 planning areas. There is little (if any at all) relationship between the planning area boundaries and the circular school boundaries,[29] and the same can be cursorily verified through inspection of the 27 electoral constituencies, 28 postal districts, and 82 postal sectors. The school district boundaries also do not appear to coincide with major roads, railways or boundaries used for other political or planning purposes.

On the whole, appears that the main concern for this paper arises from spatial effects. These will be addressed in the next section on housing models.

[26] These papers provide methodological reviews of discontinuity studies followed by recommendations for justification of the RDD in applied research.
[27] This terminology is used widely in the causal effects literature, following Rubin (1974), and in this context it refers to the effect of gaining admission priority, a result of being within the boundary, on housing price.
[28] Compound treatment effects refer to the incidence of multiple mechanisms resulting from the definition of the treatment group, such as if houses on either side of the boundary received different planning policies in addition differences in priority admission.
[29] The boundaries are shown only for schools which underwent a ballot for admissions in the past 3 years.

Figure 3.1: Map of Singapore

55 Urban Planning Area Boundaries and 187 Primary Schools, with Boundaries



3.4 Estimation Methodology

In this section I outline the three main strategies for estimating the PAV; the first two are RD estimators are the third is a matching estimator. The RD methods involve a two-stage process. The first stage is a "residualizing" approach (Lee & Lemieux, 2010, p. 331), where log house price of the i^{th} housing unit, (Y_i), is regressed on a vector of covariates:

$$Y_i = \alpha + \rho W(Y_i) + \beta X_i + \epsilon_i$$

Here X_i is a vector of housing characteristics, $W(\cdot)$ is an inverse distance weighted spatial lag operator, and ϵ_i the error term. The residuals $r_i = Y_i - \hat{Y}_i$ are saved from this estimation. The idea behind this is to first isolate the effects of covariates from housing price, enabling the treatment effect along with any underlying polynomial trends to be more discernible. Thus, the estimation utilizes the whole dataset of observations, to improve precision. Lee and Lemieux (2010) demonstrate how, as long as the housing price model is specified correctly,

the treatment effect in the second stage will be consistently estimated, and will in general be unaffected by the first stage estimation.

In the second stage, estimation is performed only on the subset of observations that are within 500 meters (0.31 mi) of a GEP School. There are three different estimation strategies that I employ: parametric RD, nonparametric RD and matching.

1. Parametric RD Estimation

This strategy involves specifying a parametric model. In the simplest case, this is equivalent to a linear model estimated on observations within a fixed bandwidth around the boundary. An issue with this strategy is that there are several key selections to be made that could affect the effect estimated. They are: the order of the polynomial, the bandwidth around the boundary for estimation and the kernel weighting scheme used. The baseline parametric model (order 0), with treatment dummy D_i, specified is:

$$r_{is} = \kappa + s_i + \phi D_i + u_{is}$$

and additional interaction terms $\sum_{j=1}^{k} \gamma_{j,1} Z_{is}^{j} + \gamma_{j,2} D_i Z_{is}^{j}$ are added for a polynomial of order k. The forcing variable, distance (Z_{is}), is specified with reference to the boundary rather than with reference to the school, and so is negative where the i^{th} housing unit is "within" the boundary, and positive if outside it. The PAV is estimated through the use of Ordinary Least Squares on the model to recover the estimate of the parameter $\hat{\phi}$. Standard errors are clustered at the school level.

2. Nonparametric RD Estimation

An alternative to a parametric model is nonparametric estimation (Hahn, Todd & van der Klaauw, 2001; Porter, 2003). Unless the relationship between housing price and distance to the boundary is exactly linear, this method will induce bias in estimates; however, this bias will be small when the sample size is large, owing to the bias convergence properties of the

estimator (Fan & Gjibels, 1992; Porter, 2003).[30] Similar to the parametric setup, the selections to be made in this context are the bandwidth and kernel type. Several "automatic" bandwidth selectors have been proposed,[31] and I use the one proposed by Imbens and Kalyanaraman (2012), which is data driven and produces reasonable figures on this dataset. As for kernel type, Cheng, Fan and Marron (1997) have shown that the theoretically-correct kernel to use for nonparametric estimation at a boundary is the triangular (edge) kernel.

3. Matching Estimation

The final model specification follows guidelines proposed by Gibbons, Machin and Silva (2013), and can be described as a spatial matching approach.

$$Y_{is} - Y_{js} = \phi(D_i - D_j) + (\epsilon_{is} - \epsilon_{js}) = \phi + (\epsilon_{is} - \epsilon_{js})$$

The underlying specification of the model is similar to the one above, except that first, each house within the boundary is matched to a "similar" observation outside the boundary, then differences in log prices are taken. Here I let i and j refer to indices of houses within and outside the boundary, and the error term to encompass unobserved spatial attributes. Differencing eliminates the terms κ and $(s_i - s_j)$ and simplifies $(D_i - D_j) = 1$. When i and j are geographically close together, $WY_{is} \approx WY_{js}$, and spatial effects are kept to a minimum.

Keele & Titiunik (2015) argues for using geographic distance as the sole basis for matching, while Gibbons, Machin & Silva (2013) proposes matching additionally on covariates. If matching is done on the basis of covariates, the term $X_{is} - X_{js}$ is also approximately 0. Thus the difference in log prices of matched pairs must be accounted for by difference in treatment status, and $\hat{\phi}$ identifies the value of priority admission.

[30] This means that the estimator proposed, in comparison to others estimators (such as the Nadaraya-Watson), achieves the fastest rate of bias reduction per unit increase in sample size.
[31] Imbens and Kalyanaraman also provide a review of the alternatives, such as the DesJardins and McCall (2008) bandwidth selector and Ludwig and Miller's (2007) cross-validation method.

4. Descriptive Statistics and Preliminary Analysis

4.1 Summary Statistics and Exploratory Data Analysis

After observational selection, the original dataset consists of 50,330 and 19,674 public and private housing transactions respectively. Table 2 in Appendix A presents some summary statistics of apartment characteristics.

In addition, I present plots of the relationship between covariates and log price in Figure 3 of Appendix A. These plots give insight as to the usefulness of the predictors in explaining house price, and how they should be entered into the linear model. The symmetry of age and years to lease expiry for public housing reflects that the two are perfectly correlated (lease expiry is set at 99 years from construction date) and so the latter should not be included in HDB price regressions. The variables time and age appear to warrant the inclusion of a quadratic term. For distance to MRT, Wong (2011) utilizes a binary variable than equals 1 if a housing unit is within a 300 meter (0.18 mi) radius; in the graph, it does appear that the relationship tapers off after awhile, so I follow this specification as well. The relationship between price and geo-coordinates will be addressed in the next section.

Figures 4.1 and 4.2 show the distribution of private and public housing prices across Singapore. Singapore's Central Business District (CBD) is located at the southern part of the island. In general, cities tend to experience radially decreasing trends away from the center. Singapore's housing landscape exhibits a similar trend, with the highest decile of both types of housing clustered around the CBD. This is also reflected in the covariate plots of log price against geographic coordinates in Appendix A, and it can be seen that the trends are more pronounced for private compared to public housing.

Along with the physical covariates, I fit several regression models to the housing price data and present the results in Tables 4A and 4B of Appendix A. The aim is to choose suitable orders for the global spatial trends. Although the addition of high order polynomials and

cross-interactions for the X- and Y- coordinate variables were all statistically significant, up till order 6, I was wary of overfitting these global trends to noise that would be better captured by local spatial weights. In the end, I chose a quadratic term in X for both housing types, a linear term in Y for HDB data, and a quadratic term in Y for private housing prices.

Figures 4.3 and 4.4 present the geographic distribution of residuals from the selected regressions (column 3 each of Tables 4A and 4B). The diagrams show that the apparent spatial clusters observed from Figures 4.1 and 4.2 cannot be fully explained on the basis of physical covariates or global spatial trends. The global Moran's I statistic, a measure of spatial autocorrelation, is 0.74 and 0.64 for public and private housing log prices respectively, using a 150m binary weights matrix. Even after controlling for covariates and global spatial trends, the Moran's I statistics computed on the residuals drop to 0.59 and 0.53 respectively. This reflects a high degree of residual spatial autocorrelation between geographically close observations, and suggests further that there are important local spatial effects influencing housing prices, which motivates the inclusion of spatially weighted variables as regressors.

Figure 4.1: Distribution of HDB Prices across Singapore

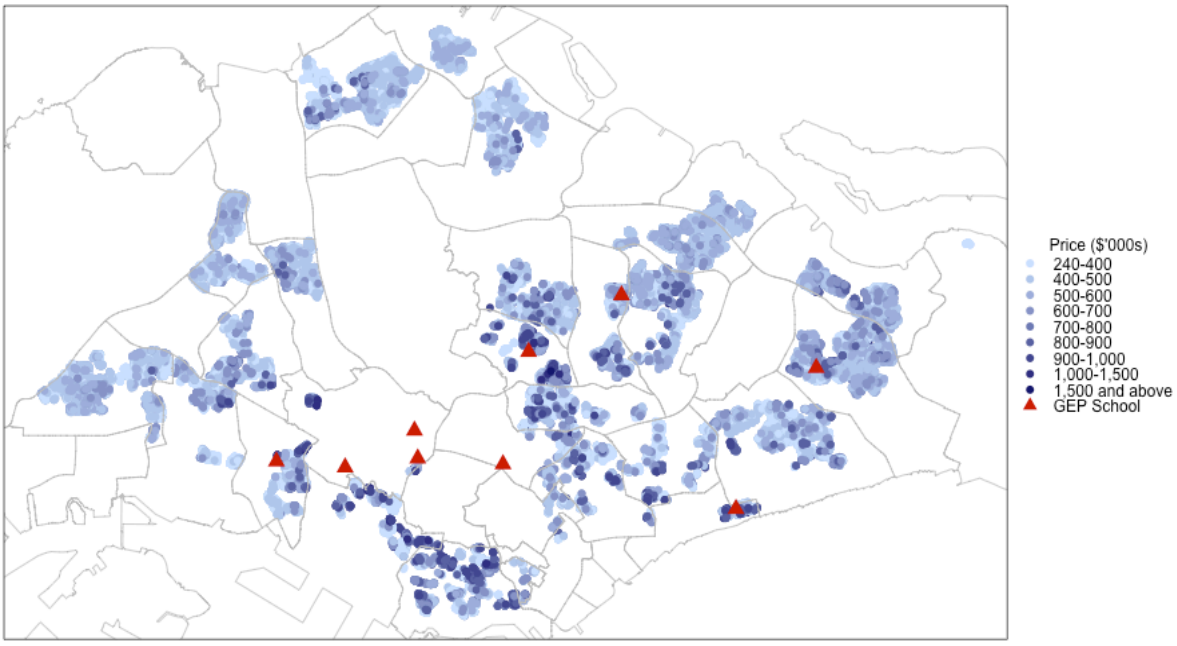


Figure 4.2: Distribution of Private Housing Prices across Singapore

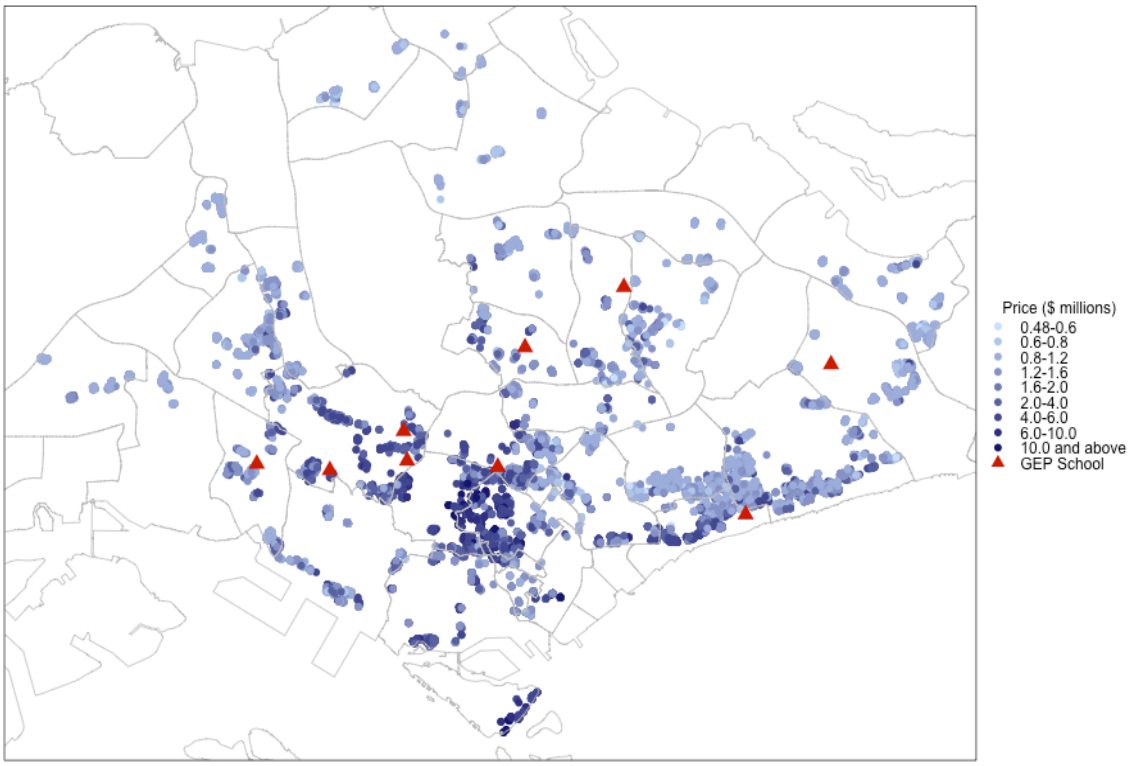


Figure 4.3: HDB Price Residuals from Trend Surface Regression

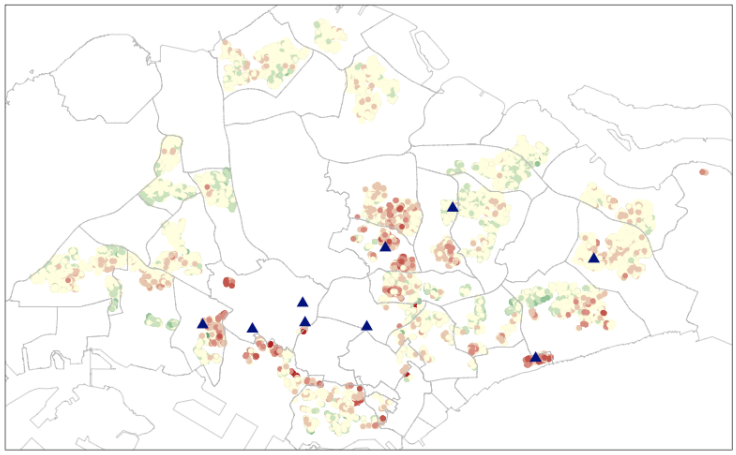


Figure 4.4: Private Property Price Residuals from Trend Surface Regression

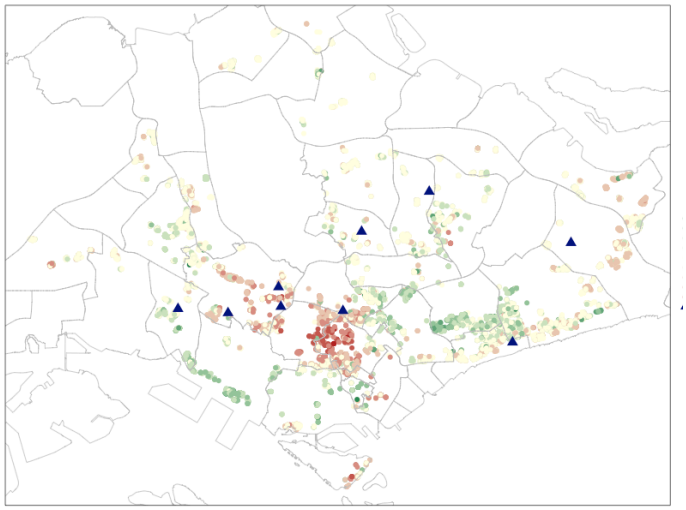


Figure 4.5. Test for Spatial Autocorrelation: Moran's I Coefficients*

	Public Housing Prices					Private Housing Prices				
Weights Matrix	*100m*	*150m*	*200m*	*250m*	*Avg*	*100m*	*150m*	*200m*	*250m*	*Avg*
Spatial Lag										
No Lag	0.754	0.736	0.708	0.681	0.72	0.627	0.644	0.648	0.657	0.64
100m IDW Lag	0.354	0.413	0.396	0.377	0.39	0.138	0.178	0.194	0.201	0.18
150m IDW Lag	0.376	0.368	0.365	0.35	0.36	0.178	0.138	0.172	0.188	0.17
200m IDW Lag	0.401	0.392	0.358	0.344	0.37	0.19	0.15	0.139	0.166	0.16
100m Binary Lag	0.356	0.414	0.398	0.379	0.39	0.134	0.176	0.193	0.2	0.18
150m Binary Lag	0.406	0.376	0.371	0.357	0.38	0.205	0.139	0.176	0.194	0.18
200m Binary Lag	0.447	0.419	0.372	0.355	0.40	0.231	0.166	0.138	0.167	0.18

*** All Moran's I coefficients are significant at the 0.1% significance level, using a permutations test**

4.2 Choosing Spatial Lag Order

The next step is to choose an appropriate order for the spatial lag for the final global model specification. I construct spatially lagged variables of orders 100m, 150m and 200m, (328ft., 492 ft. and 656 ft.) using binary and inverse distance weights (IDW). A small percentage of observations have no neighbors within given bandwidths, and I impute a 12-nearest-neighbor-based spatial lag value. I also create K-nearest-neighbors (KNN) lags, which do not suffer from the problem of missing values, choosing K = 5, 10, 12, 15, and 20.

In choosing the spatial model, I consider two things: the amount of residual spatial autocorrelation in the residuals, as measured by Moran's I, as well as the significance of the spatial lag variables in the regression. Tables 4C and 4D in Appendix A present the coefficients from the spatial regressions, with only the three top-performing models from each lag type (Distance-weighted and KNN) shown.

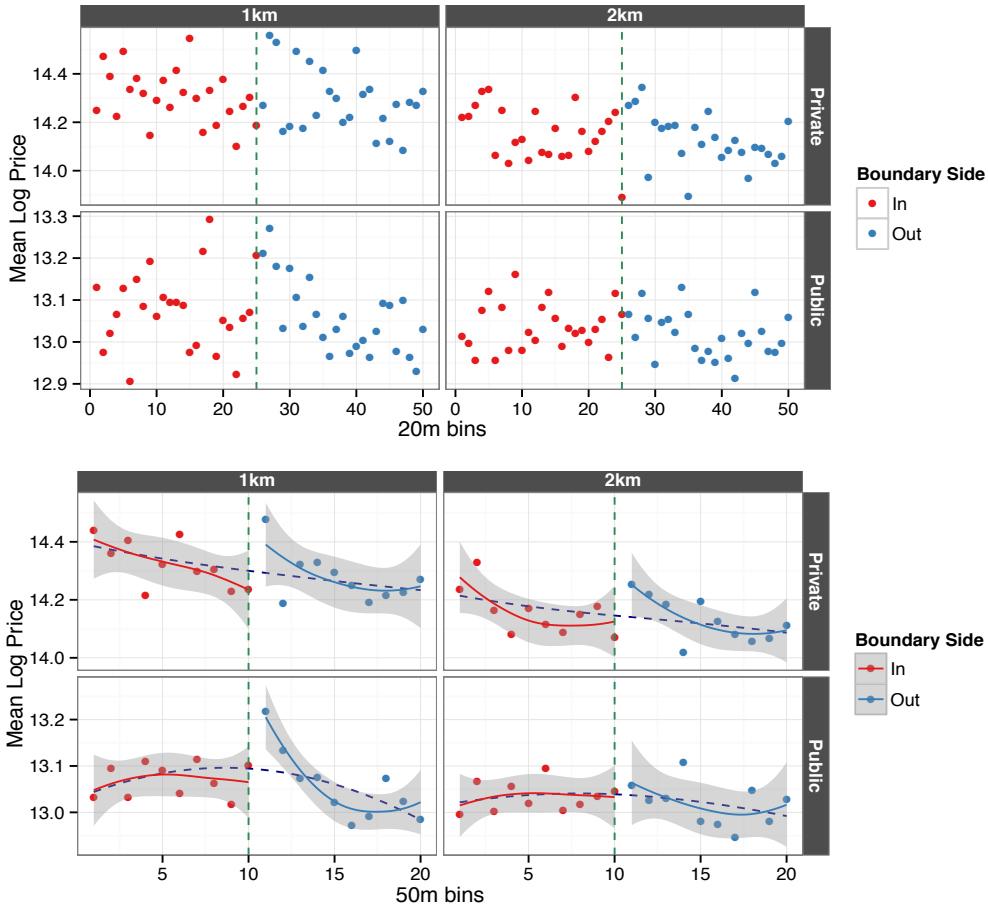
The final lags chosen were IDW 150m and KNN-10 for public housing, and IDW 200m and KNN-15 for private housing. The level of spatial autocorrelation is reduced significantly, by about 40% and 70% respectively, for public and private housing prices. All coefficients are highly practically and statistically significant, and there is little difference in performances across models within each type.

The residuals from the respective regressions are saved, along with those from the aspatial case for comparison, for use in the next stage of analysis. At this stage, I also subset out the observations within 500 meters of the school boundaries, which is the maximum distance that avoids overlapping observation sets between boundaries. I run all subsequent analyses on four separate datasets, HDB and private housing at the 1km and 2km boundaries, which I refer to using the abbreviations H1, H2, P1 and P2 henceforth.

4.3 Preliminary Justification of the RDD

Prior to conducting the actual discontinuity estimation it is helpful to examine plots to check for a discontinuity in the case of the conditional regression function and for continuity in the case of other variables. This involves dividing observations into "distance bins" of 20m (66 ft.) and 50m (164 ft.), taking the average of each bin for a given variable, and plotting the results by bin. What this does is to give a visual representation of how that variable, for instance price, evolves with respect to distance, towards and away from the boundary. I first present binned plots of the prices in Figure 4.6 (inset), with accompanying LOESS fits in the 50m case, to check for any apparent discontinuities.

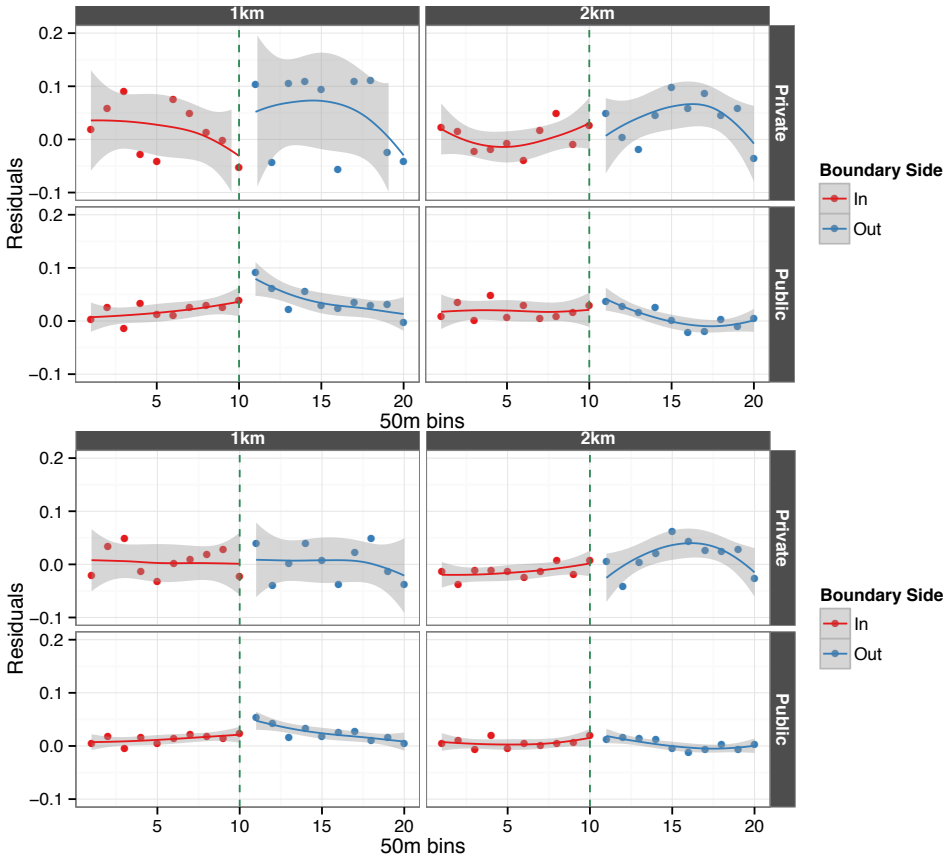
Figure 4.6. 20m and 50m Binned Plots of Log Housing Prices



Evidence for a possible discontinuity points us in the other direction, with housing prices just outside the 1km boundary (the blue points) being worth slightly more than those slightly inside. The error bars of the LOESS fits mostly coincide at the boundary, with the

exception being the H1 case (bottom left). In Figure 4.7 (inset), I repeat the process but this time show binned residuals from the preliminary regressions instead.

Figure 4.7. 50m Binned Plots of Residuals from Linear Regression (above) and Spatial Regression with KNN Lag (below)

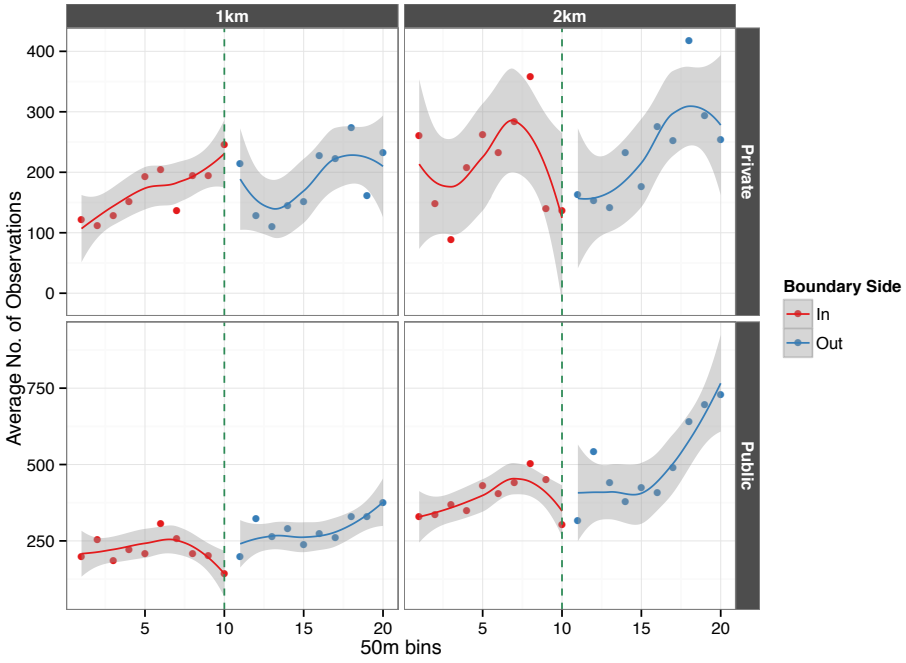


The above plots were produced using the same y-axis scale. It can be observed that the P1 data are the noisiest, and that the inclusion of the spatial lag reduces this noise by a significant amount, enabling the underlying trends to be more discernible. Again, there does not appear to be visible evidence of a discontinuity, except in the H1 case (bottom left), which has shrunk from a magnitude of around 0.1 to about 0.03. Of course, these plots are not proof of a null effect yet, since the estimation thus far has not controlled for schools, and the frequency of observations are not yet taken into account here.

Other recommended plots are binned averages of the covariates, which are provided in Figure 5 of Appendix A. By and large, most of the covariates appear to evolve smoothly across the boundary. The exceptions are floor area and floor level, for H1 and H2; housing

units just outside the boundaries appear to be slightly more valuable in terms of having more of both variables. This might partially explain the discontinuity in the H1 data, but even after controlling for these covariates the discontinuity still remains.

Figure 4.8. Observation Density around the Boundaries (50m bins)

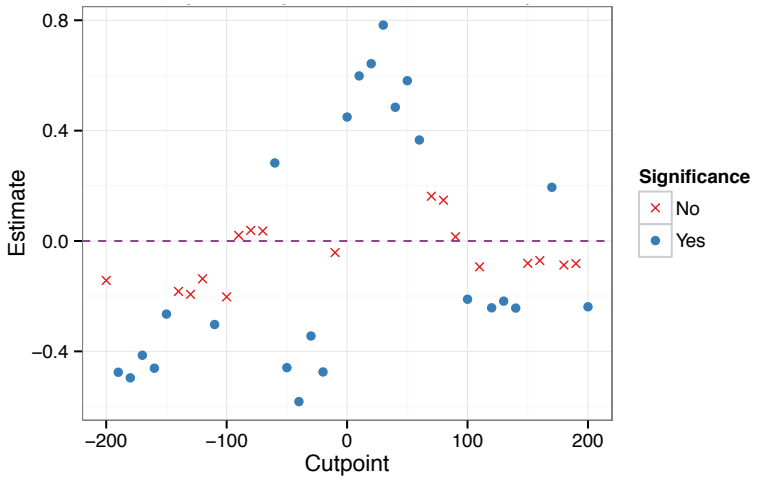


Of interest also is the density plot of the running variable, in this case distance from the boundary, which is presented in Figure 4.8 (inset). Theoretically, a discontinuity in the density of housing at the boundary would suggest manipulation of the running variable. As distance from school increases, there is more land area and we would expect a higher frequency of observations, which is confirmed by the plot in all four cases. There also appears to be a small discontinuity in the H1 and P1 cases, and this is seconded by the McCrary's (2008) density test results shown in Figure 5 of Appendix A.

Although the test rejects the hypothesis of no sorting at the 5% significance level for H1 and P1 data for all binwidths, in practice this is not a big concern for several reasons. First, if there were manipulation, we would expect the density of observations within the boundary to be higher than outside it, so the discontinuity is unlikely to reflect foul play. The test results are not stable even for the H2 and P2 cases; moreover, running the test across

different cutpoints at 10m intervals produces a rejection of the null hypothesis 59% of the time (Figure 4.9, inset). This indicates that the housing landscape is generally discontinuous by nature, as we would expect since most sales are from apartments in high-rise blocks. Moreover, given the strict planning regulations governing residential development in Singapore, we would not expect supply-side interferences to be probable. Finally, given that apartment blocks are generally immovable, it is unlikely that there other forms of manipulation are even possible.

Figure 4.9. McCrary Density Test at different "imaginary" cutpoints



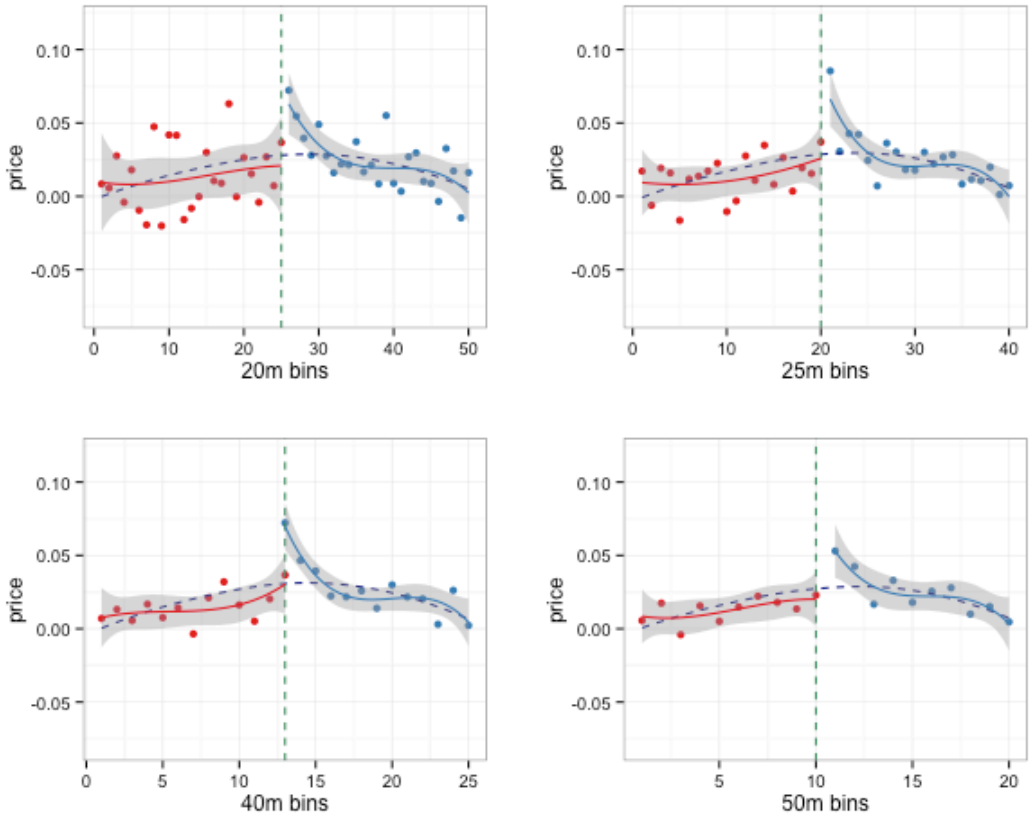
In sum, a few of the assumptions traditionally required for establishing the credibility of the discontinuity design are in question. Upon closer contextual scrutiny, the apparent violation of these assumptions does little to logically challenge the interpretability of any discontinuity found as a measure of the PAV.

5. Results from Individual Models

5.1 Parametric RD Results

To run a parametric discontinuity model, it is necessary to specify the order of the polynomial and the bandwidth. The rationale for including a polynomial is to allow for the running variable, in this case distance to the school, to influence prices in a non-monotonic fashion. Given the high density of housing and schools in Singapore, and the circularity of boundaries, there is little reason to believe that prices should evolve uniformly in a radial fashion from school locations, especially as school boundaries overlap with each other. Nevertheless, due to the inability of the spatial model to capture all of the spatial dependence, especially in the public housing data, I still fit several low-order polynomials to the data. One feature of the data is that there is a modest bandwidth (500 meters) of observations around the boundaries, thus the inclusion of a high-order polynomial is likely to be unjustified.

Figure 5.1. Fitting a polynomial of order 3, H1 dataset, bandwidth 500m



I use two methods to choose bandwidths and orders for the parametric model. The first method is a manual approach where I attempt to fit polynomials "by eye" using visualizations for each of the datasets. In Figure 5.1 (inset), a polynomial of order 3 is chosen because the polynomial fit is stable across all 4 bin-width specifications, and visually "fits" the data as well. Where the data is noisy, such as in Figure 5.2 (inset) for the P1 dataset, a linear fit is used instead, especially where the cubic fit looks less appropriate with the use of finer bin-widths, such as in Figure 5.3 (inset).

Figure 5.2. Fitting a polynomial of order 1, P1 dataset, bandwidth 500m

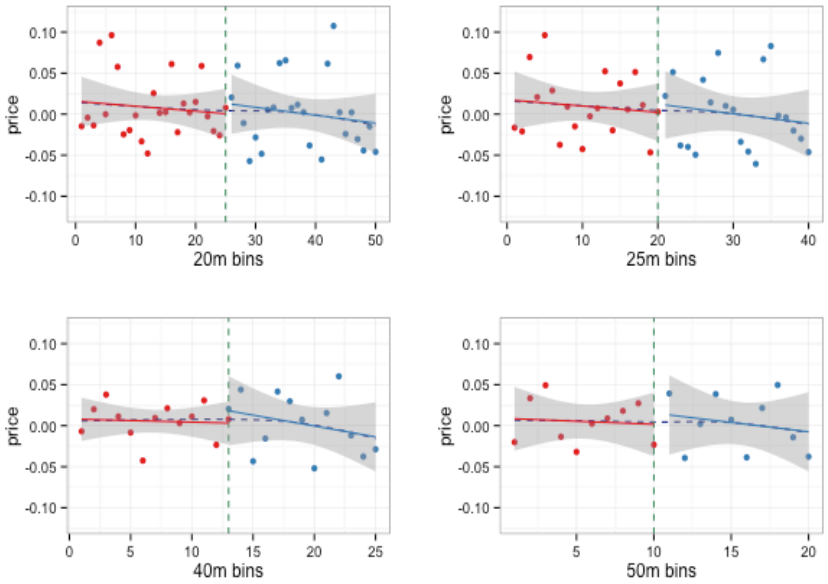
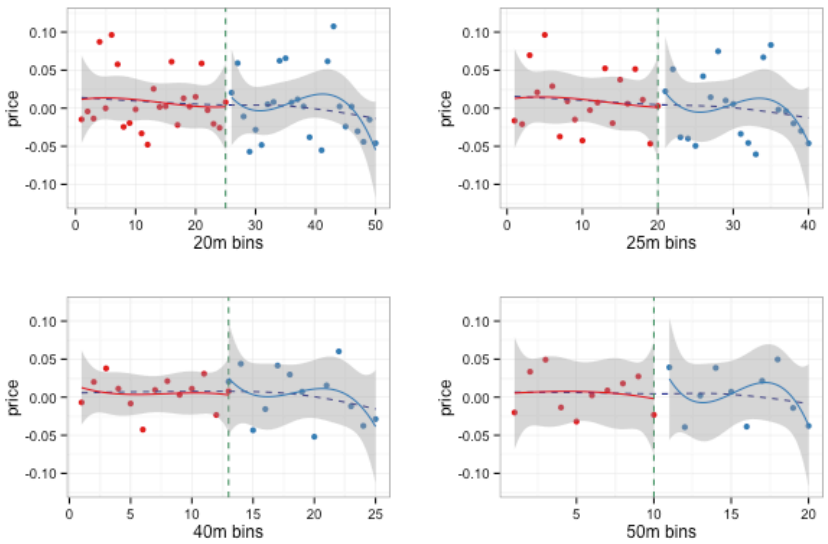


Figure 5.3. Fitting a polynomial of order 3, P1 dataset, bandwidth 500m



The final orders for the polynomials for the different bandwidths are shown below in Table 5.4 (inset). As aforementioned, lower order polynomials tend to be chosen for narrower bandwidths. The second method is to use F-tests to test for the significance of the additional polynomial terms. For instance, if a d-ordered polynomial is currently fit to the model and the addition of the $(d+1)^{th}$ order terms does not produce a significant change in the fit, then the F-test fails to reject the null hypothesis and I fit a polynomial of order *d*. Otherwise, I repeat the cycle for the $(d+1)^{th}$ polynomial and so on. Expectedly, the maximum order required never exceeds 3. The estimation results are presented in Figures 5.5 and 5.6.

Table 5.4 Polynomial Fit Selection Results

	Bandwidth				
Visual Polynomial Fit*	*200m*	*250m*	*300m*	*400m*	*500m*
Public Housing, 1km	1/2	1/2	1/2	1/2	1/3
Public Housing, 2km	0/1	1/2	1/2	1/3	1/3
Private Housing, 1km	0/1	0/1	0/1	0	0
Private Housing, 2km	0/1	0/1	1	1	2/3

Repeated F-Tests	*200m*	*250m*	*300m*	*400m*	*500m*
Public Housing, 1km	1	1	3	2	1
Public Housing, 2km	0	1	1	2	2
Private Housing, 1km	0	0	0	0	0
Private Housing, 2km	0	0	3	1	1

Notes:
- Numbers represent the order of the polynomial selected
- * Slashes (/) indicate where more than one polynomial produced reasonable results

Figure 5.5. Results from parametric models using a visual fit

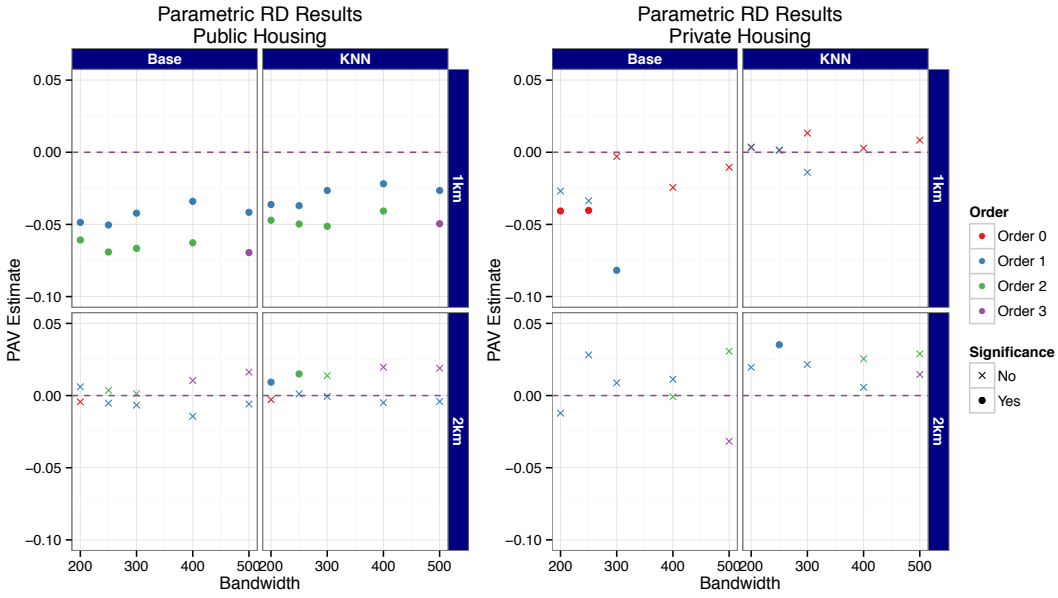
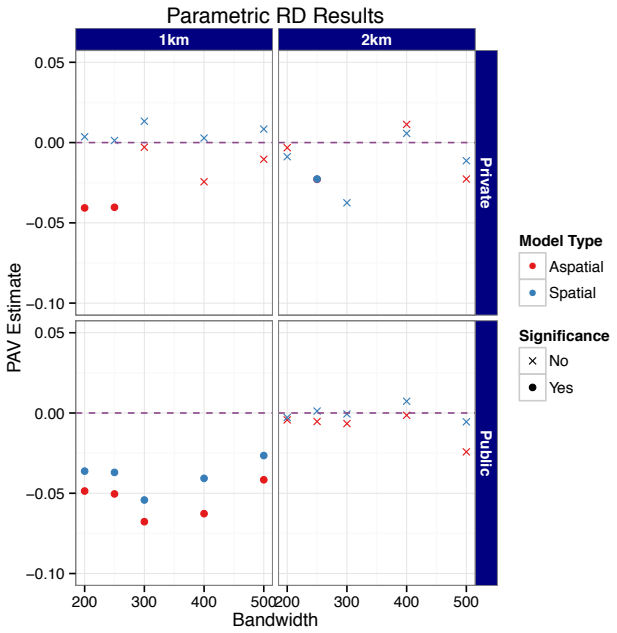


Figure 5.6. Results from parametric models chosen with repeated F-tests



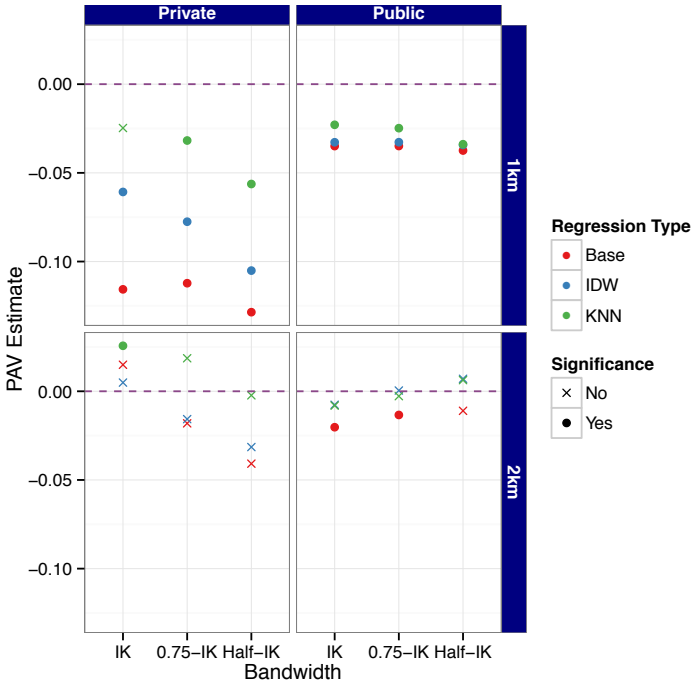
Finally, it is possible to start answering the research questions laid out at the beginning of this paper. The exploratory analysis so far has been consistent with the lack of a discontinuity, with the sole exception being the H1 data, which show a discontinuity in the opposite direction. This means that houses right outside the 1km boundary are actually valued more than those just inside. Although this was foreseeable given the binned plots presented in Figure 4.6, the finding is, practically, surprising. As for the others, the H2 estimates are quite convincingly around zero, while the other two cases are unclear.

As for the second question on the treatment of space, the spatial model estimates appear to move in tandem with the aspatial estimates most of the time, when we alter the order or the bandwidth. If controlling for space using a spatial lag were crucial to model validity, then we would expect to see the spatial estimates behaving more stably than aspatial estimates. Preliminarily, this is not immediately obvious from the above plots, and as the fits were made on the basis of the residuals from the regression using the KNN lag, we would expect the spatial model to produce more stable estimates anyway.

5.2 Non-Parametric RD Results

Next I present the results from the non-parametric RD model. The Imbens and Kalyanaraman (2012) bandwidth (IK bandwidth) selector provides reasonable and believable estimates of the bandwidth to be used, with the average bandwidth across the four datasets and two specifications at 362m. In addition to the results using the IK bandwidth, I also use 0.5 and 0.75 times of the IK bandwidth. This is in accordance with Keele & Titiunik (2015) and Calonico, Cattaneo & Titiunik (2013) who find that in the context of space the IK bandwidth is often "too wide", and instead recommend "under-smoothing" by using smaller, albeit arbitrary, distances. The results are presented in Figure 5.7 (inset).

Figure 5.7. Non-Parametric RD Results



Again we find stable and negative estimates for the H1 data, all of which are relatively stable and invariant to the bandwidth. The 2km cases are once again come close to reflecting a lack of an effect, but with some systematic variation across the choice of bandwidth. This may be indicative of entire a set of outliers or some residual spatial trends yet to be controlled for. The most interesting results come from the P1 data, which now reflect a very large negative PAV, and which differ largely based on the type of lag used. The unweighted model

produces the largest magnitude, in excess of 10% of house price, while the IDW-lagged model followed by the KNN-lagged model each cut away an additional 3-5% of the magnitude. Given the large standard errors in the binned plots, this should not be altogether surprising; however, the next problem concerns how to interpret the bizarre results. I discuss these results in full in Section 6.

5.3 Exploratory Matching Analysis[32]

In matching, the key choice is the set of covariates to match observations on. Keele and Titiunik (2015) recommend matching observations on geographic distance; my first specification is therefore to match on coordinates, which is almost equivalent.[33] I also experiment with matching purely on covariates and matching both on coordinates and covariates (Gibbons, Machin & Silva, 2013). The balance statistics that result from the matching procedure (using all observations) are displayed in full in Table 7 of Appendix A, while a condensed summary is presented in Table 5.7 (inset). The three statistics measure equality in variances, means, and distributions respectively. So far, all the physical covariates

Table 5.7: Summary of Matching Performance Statistics

		P-Values				Min/Max Variance Ratio			
Variable Type	**Matching Type**	**H1**	**H2**	**P1**	**P2**	**H1**	**H2**	**P1**	**P2**
Flat Type*	Unmatched	4/7	3/7	0/4	0/4	0.11 / 2.88	0.78 / 1.5	0.04 / 1.55	0.57 / 1.27
	Coordinates	1/7	2/7	0/4	1/4	0.11 / 141.16	0.49 / 22.76	0.04 / 2.15	0.86 / 1.13
	Covariates	5/7	7/7	4/4	4/4	1.00 / 1.03	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	Coordinates and Covariates	5/7	7/7	4/4	4/4	0.98 / 1.03	1.00 / 1.00	1.00 / 1.00	1.00 / 1.01
Other Physical Attributes**	Unmatched	3/13	7/13	7/13	6/13	0.60 / 1.26	0.71 / 1.05	0.69 / 1.04	0.77 / 1.33
	Coordinates	0/13	0/13	0/13	3/13	0.43 / 1.07	0.60 / 1.18	0.49 / 1.68	0.53 / 1.68
	Covariates	6/13	7/13	7/13	3/13	1.00 / 1.26	1.00 / 1.16	0.99 / 1.36	1.01 / 1.21
	Coordinates and Covariates	5/13	6/13	4/13	2/13	1.01 / 1.19	1.03 / 1.17	1.00 / 1.35	1.03 / 1.34

- P-Values:
No. of P-Values above 0.05 / Total No. of P-Values (Higher numbers are better)
- Min/Max Variance Ratio:
Minimum nonzero variance ratio / Maximum variance ratio (Values closer to 1 are better)

NOTES:
*** 7 types of Public Housing and 5 types of Private Housing (only 4 types within 500m of schools)**
**** 7 Attributes measured: age, squared age, floor area, floor level, time, squared time, MRT proximity**
Red: Worsened statistic (over unmatched case)
Blue: Improved statistic (over unmatched case)

[32] To implement matching using R, I use the "Matching" package, details of which are found in Sekhon (2011).
[33] This is equivalent to matching on the basis of Manhattan distance, whereas geographic distance implies Euclidean distance. In practice this should not alter the results by a great deal.

have exhibited significant covariation with price. As we would like to minimize their effects as far as possible, the approach that achieves the best balance will be taken as the benchmark.

On the whole, matching purely on geographic distance does not do very well in terms of covariate balance, often worsening the balance instead of improving it. This is somewhat of a concern, since the RD approach relies mainly on geographic distance as a basis for comparison of housing types, through bandwidth choice.

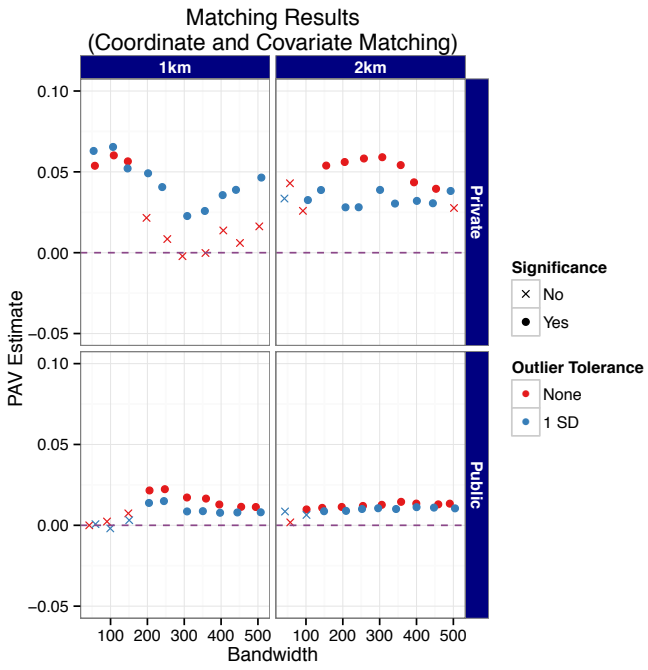
In addition, between matching on covariates and matching on both covariates and coordinates, both strategies perform similarly, with matching on covariates performing slightly better in terms of P-values across the board. However, the problem with matching on covariates is that the matching algorithm often pairs up observations that are located within different school boundaries, and the discrepancies in neighborhood quality and locational value mean we would want to minimize this occurrence as far as possible. For example, using the H1 dataset, 324 observations were matched with observations more than 2km away for matching on covariates while the number was 196 for matching additionally on coordinates. In conclusion, I choose a matching approach with matching on covariates and coordinates as the benchmark.

5.4 Matching Results

Figure 5.8 displays the results for the spatial matching estimator. Immediately, two features stand out: first, the large negative PAVs for H1 and P1 data found by the parametric and nonparametric RD models respectively have disappeared, being replaced by mostly positive estimates, most of which are significant as well. Second, the spatial matching estimates are not immune from bandwidth selection effects, although this is mainly for private housing prices.

The fluctuations in the estimates may be affected by the presence of outlier

Figure 5.8. Spatial Matching Results



observations. Having already removed apartments with more than 300 square meter floor area rather arbitrarily, I was wary of discarding even more observations. Instead, the matching algorithm allowed enforcement of observations with covariate values beyond a certain threshold into becoming ineligible as a match for treatment observations. To refine the matching on flat type and coordinates without eliminating too many useful candidates, I set this threshold to one standard deviation of the observation's covariate values. This has the effect of stabilizing the estimates, reducing the fluctuations substantially. In the case of the P2 data, the relative invariance of the results to this change hints at the possibility of these fluctuations being driven by outlier values.

The results from the spatial matching model are particularly interesting because the "obvious" negative PAVs discernible from the binned raw prices in Figure 4.6 have not only disappeared but been reversed by this model, where a large majority of the estimates fall into the positive domain. In addition, the estimates for the H1 and H2 data appear to be fairly stable and well-behaved.

6. Discussion

Having presented the results from the individual models separately, I now attempt to reconcile all the results thus far by comparing the results across the different methodologies and specifications. The first step is to sort out how the RD and matching approaches size up next to each other. This will enable a case to be made for choosing one approach over the other. Next, I address the use of spatial lags and model specification. Finally, we arrive full circle at the question that motivated this entire paper – the value of the PAV.

6.1 Interpretation of the Residual

The larger question at hand is whether the value of the PAV can be measured by a comparison of housing prices within a small bandwidth around the boundary, an approach used by Black (1999), and also in Wong's (2011) paper on an earlier Singapore housing dataset. I would argue that even within a small bandwidth, houses on either side may not be directly comparable due to development-specific effects, and omitted variables bias. The main evidence from the results so far are the lack of balance in the matching summary (Tables 7A and 7B of Appendix A), the presence of a substantial Moran's I in the residuals, (Tables 4C and 4D of Appendix A) and the inconsistency of results across the different specifications. In this section I discuss possible results for the pitfalls experienced by the RD, and why the spatial matching approach performs better.

6.1.1. Housing Characteristics and their Implications

Many blocks of housing units in Singapore are constructed as part of larger clusters or developments, especially in the case of condominium units for private housing, but increasingly so as well for public housing, where the HDB engages developers to develop a cluster of blocks at the same time. Within a development, its apartment blocks tend to be in very close proximity to one another, meaning that it is more often the case that all blocks are

on the same side of the boundary, than to be split on either side. On this basis alone, a simple comparison without controlling for attributes in any way is likely to be measuring differences in fixed effects specific to each development, which could be very large in comparison to a hypothetical PAV. These fixed effects may include very localized factors (e.g., high floors commanding an extra premium as a result of the development being the tallest in the region or near the sea), or may instead be associated with unobservable factors (e.g., developer prestige or amenities and facilities), that apply to all observations in the same development, and the PAV is nested within this array of possible omitted variables.

It is not feasible to directly control for these fixed effects through specifying a separate coefficient for every development or housing block. In the latter case, all units within a block either do or do not have priority admission, so controlling for block effects would difference away the PAV. Except in the case where developments happen to be split across the boundary, the same result would occur with including development fixed effects directly.

The other feature of the housing landscape in Singapore is its discontinuous observation density, which arises due to the high-rise nature of housing. This means that sample size changes discontinuously even where bandwidths evolve smoothly. This will in general produce sharp changes in estimates where subsets of observations to estimate on are defined using geographic measures.

Combining these two features results in the propensity for covariate imbalance, which means that houses on either side of the boundary are not directly comparable. This issue remains is not solved by increasing the bandwidth either, owing to the third feature of housing in Singapore – the heterogeneity and non-monotonic evolution of land value across neighborhoods as evidenced by the maps in Figures 4.1 and 4.2 (inset). As an example, table 6.1 presents balance statistics of unmatched H1 data at different bandwidths of 50m, 200m and 400m. The numbers show that increasing the bandwidth does not necessarily improve

balance. Most P-values for the T-tests and KS-tests worsen when the bandwidth changes from 50m to 400m, while the variance ratio is outstandingly large for the Premium Apartment flat type. At the same time, the large P-values for the T-test in the 50m case does not necessarily reflect balance either; they may instead be a result of a small sample size (342) with insufficient power to reject the null hypothesis. In fact, most of the variance ratios are worse for the 50m bandwidth compared to the 200m and even the 400m.

Table 6.1. Balance Statistics for different Bandwidths

Unmatched Data	**Bandwidth**								
	50m			*200m*			*400m*		
Variable	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**
Age	0.90	0.28	0.27	0.91	0.00	0.00	0.82	0.32	0.00
Age (Squared)	0.94	0.19	0.27	1.00	0.00	0.00	0.90	0.03	0.00
Square Area	1.63	0.27	0.00	1.26	0.65	0.00	1.27	0.00	0.00
Floor Level	0.65	0.01	0.01	0.56	0.00	0.04	0.57	0.00	0.00
Time	1.12	0.82	0.67	1.07	0.02	0.02	1.05	0.01	0.00
Time (Squared)	1.17	0.64	0.67	1.12	0.01	0.02	1.08	0.01	0.00
MRT Proximity (Binary)	0.12	0.00	NA	1.32	0.03	NA	1.54	0.00	NA
Type (Apartment)	2.18	0.03	NA	1.88	0.00	NA	2.40	0.00	NA
Type (Improved)	1.27	0.08	NA	0.97	0.57	NA	1.09	0.03	NA
Type (Model A)	1.06	0.37	NA	0.95	0.02	NA	0.96	0.05	NA
Type (New Generation)	0.65	0.03	NA	1.13	0.10	NA	1.02	0.65	NA
Type (Premium Apartment)	NA	NA	NA	56.23	0.00	NA	29.26	0.00	NA
Type (Simplified)	0.75	0.31	NA	0.90	0.07	NA	0.43	0.00	NA
Type (Standard)	0.00	0.00	NA	0.11	0.00	NA	0.14	0.00	NA
School (CHS)	1.14	0.45	NA	0.94	0.43	NA	0.68	0.00	NA
School (HPPS)	1.43	0.30	NA	0.39	0.00	NA	0.68	0.00	NA
School (NHPS)	0.53	0.00	NA	1.00	0.96	NA	1.00	0.96	NA
School (RS)	1.24	0.19	NA	0.92	0.11	NA	1.13	0.00	NA
School (SHPS)	1.32	0.03	NA	1.16	0.00	NA	1.15	0.00	NA

- **Var Ratio**: The ratio of the mean treatment value over the mean control value. 1 indicates perfect balance
- **T-Test**: The t-test of differences in means
- **KS Test**: The Kolmogorov-Smirnov test for difference of distributions; does not return a P-Value for binary variables

6.1.2. Covariate Control

Next, I turn to the issue of covariate "control." I would further argue that a traditional linear model is unable to fully characterize and model these housing development fixed effects. The main culprit here is space once again. In an ideal world, with infinite sample size, separate fixed effects would be specified for each location. With the linear model and finite sample size, the number of dimensions to consider, as well as possible interactions, is too

high.[34] Although I have included school fixed effects in my estimation, it would not be unreasonable to surmise separate slopes for each location as well. For instance, floor level would be more valuable in locations with a good view, such as near the sea or in a low-rise neighborhood; floor area would be likelier to incur a different premium in a central area as compared to at the fringes of the city.

If the aim were to predict the price of an individual housing unit, then the linear model may do a reasonable job up to this point. However, what are of interest are the residuals from the regression. Controlling for covariates in a linear regression type model postulates that the same relationship holds across all observations, regardless of development effects or space. A given housing unit that has been over-predicted due to omitted development-specific variables are likely to be surrounded by other units that are similarly over-predicted. This would produce positively correlated residuals and a substantial residual Moran's I statistic (as reflected in Tables 4C and 4D of Appendix A).

Another pressing issue is that of outliers, observations that have covariate values close to the end of the range, and for which the linear model does not fit as well. Moreover, linear models tend to be sensitive to such observations, and if the fit is even slightly compromised, this produces residuals with larger magnitude for the bulk of the observations even in the mid-range of the covariates, which would worsen the correlation between residuals amongst similar housing units.

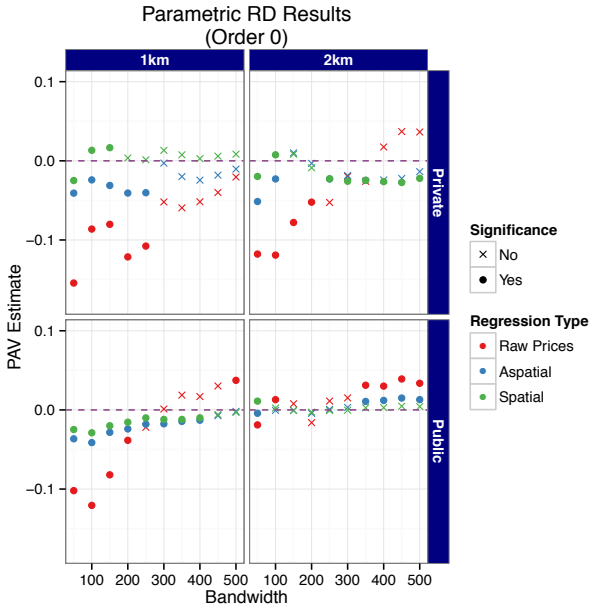
In sum, even after controlling for covariates in a linear regression type estimation, the residuals are a possible combination of bias from improper modeling of the covariate relationship, which could come from omitted interactions with space, or outliers, and finally from the PAV.

[34] It is technically possible to increase the sample size by taking a larger sample, extending the time window back several years. There then arises the problem of space-time interaction, which is best left out of the picture due to the active intervention on part of the Urban Redevelopment Authority in the housing market in the last decade in cooling property prices.

6.2 Regression Discontinuity Models

Figure 6.2 (inset) illustrates this problem within the RD estimation. Here I use a polynomial of order 0, essentially a comparison of average residual values within the bandwidth on either side. The most striking feature of the plots is the instability of the estimates on raw prices. Controlling for covariates and including a spatial lag in the regression tempers the volatility partially, but there are still residual spatial trends that cause both the spatial and aspatial estimates to fluctuate between the positive and negative domain, and in and out of statistical significance. Stability in estimates itself is not confirmation that what we have found is the true value of the PAV, but rather seems to be a precondition for making such an inference.

Figure 6.2. Regression Discontinuity Estimates over different Bandwidths



In addition, we need to consider whether the alternative explanations can be ruled out. It should also be fairly clear that neither varying the type of kernel used, varying the bandwidth used, nor using nonparametric estimation will not overcome the problems of space-covariate interaction or omitted development fixed effects. The first two only change

the relative importance and types of observations that are included in the estimation, which may actually worsen the problem depending on the context.[35] Likewise, nonparametric RD estimation, which achieves a local fit, will also be more biased if it adapts its estimates more closely to a large cluster of residuals close to the boundary.

6.3 Matching Models

In this context, the estimates based on coordinate matching (Keele & Titiunik, 2015) are inappropriate, given the shape of the boundaries. As recommended by Sekhon (2011),[36] observations are repeated in matching, which means that the control group is restricted to the observations right outside the boundaries. Increasing the bandwidth of observations simply ropes in more observations further from the boundary, which, as bandwidth increases, may be less comparable to the selected control group. Moreover, with small bandwidths, the importance of the control group is very high in the RD estimation; as the bandwidth increases, more observations further from the boundary are included, diluting the weightage carried by this group. Thus, a matching approach based solely on coordinates is equivalent to the RD estimator with order 0 and a uniform kernel for a narrow bandwidth.

Moving on to the coordinate and covariate matching estimates (Gibbons, Machin, & Silva, 2013), we see that this has a few desirable advantages over the RD estimates. First, with correlated residuals, the RD estimator will be sensitive to the density of transactions across the landscape. In contrast, matching with repeats allows each observation to be paired to an appropriate counterfactual. Consequently, in a scenario where 100 small negative residuals exist inside and only 1 large negative residual and some positive ones outside the boundary, the RD estimator would be likelier to conclude that there is a statistically

[35] For instance, a triangular kernel would amplify the bias if the observations near the boundary had particular large development fixed effects not related to the PAV, while enlarging the bandwidth would achieve the same if observations away from the boundary had a higher variance of these fixed effects.
[36] Without allowing repeat observations, the order in which treatment observations are assigned matches matters. Thus, disallowing repeats will generally induce further bias in the matching estimation.

significant negative effect of being within the boundary. However, the matching estimator may match all those within to the one large negative one outside, on the basis of covariate comparability, and produce a positive effect instead. Hence, the matching estimator is relatively more robust to discrepancies in transactional density.

The matching estimator also eliminates several of the alternative hypotheses proposed in Section 6.1. Regarding space-covariate interaction effects, a linear model is required to generalize to every possible combination of covariates and space, while matching only requires the specification of a similarity measure. As long as the similarity measure performs well, this should keep the aforementioned biases down to a reasonably low level.

As an example, I propose that the combination of age and school provide a reasonable proxy for housing estate, without differencing away the PAV. All units within the same block have the same age, and usually this extends to comparable neighboring blocks as well. However, two 10-year-old blocks in different estates would make a bad comparison, which is why the addition of school as a matching covariate is a powerful step-up from the linear model. This is illustrated in Table 6.3 (inset, next page), where it can be seen that only a few blocks correspond to each given age. If they happen to be across the boundary, such as in the case of the bottom panel, then they present the ideal scenario for matching to occur.[37]

Moreover, observations for which no comparable observations exist can be dropped, which would resolve the problem of having to model outliers. This is illustrated by adjusting the tolerance level in the algorithm,[38] which is shown in Figure 6.4 (inset, next page). Leaving aside the base case which did used all observations, it is apparent that these three other specifications are generally quite similar and consistent, with the most minimal of fluctuations across all models and specifications for the P1 and P2 data.

[37] For HDB blocks, the first two digits represent the postal sector while the third is the postal code. Blocks sharing the first three digits are in general very close to each other, possibly within a radius of 100-200m.
[38] Tolerance levels from 1.5 SD to 2 SD were found to produce exactly the same result as the 1.25 SD case.

Table 6.3. Property age as a proxy for housing development

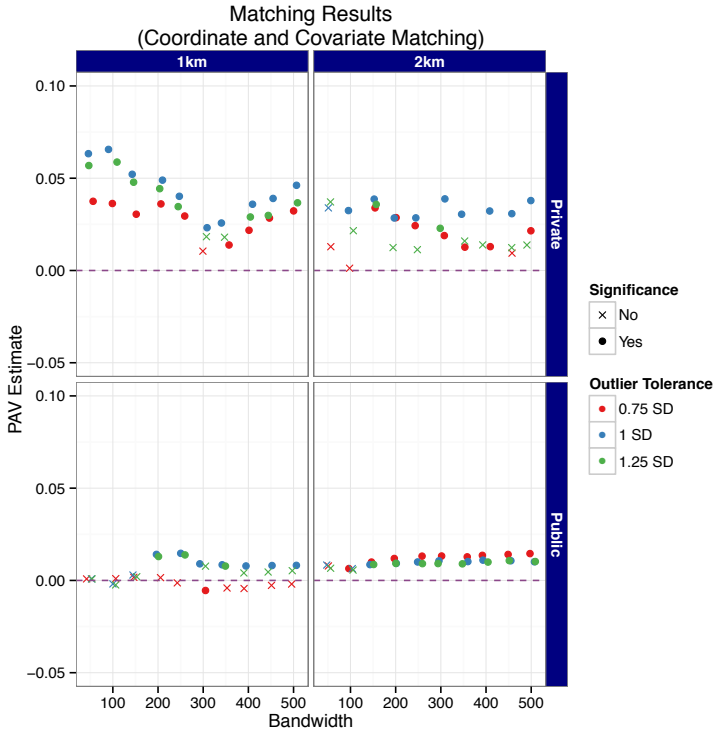
H1	Locational Indicators				
Property Age	*N*	*Postcodes (830)*	*MRT (7)*	*Schools (5)*	*Flat Types (7)*
2	53	9	3	2	2
6	69	8	1	1	1
8	71	6	2	2	2
10	49	6	1	1	2
11	112	17	3	2	2

	Flat Type		
Age = 11 School = RS	Improved	Model A	Postcodes
Within	12	19	531979, 532979, 533978, 533979, 536978
Out	16	36	531980, 531981, 532980, 532981, 533980, 533981, 536980, 537981

NOTES:

- Locational Indicators refer to the number of unique values taken by all units of a given age
- Number of brackets indicates total number of values for all observations in the H1 dataset

Figure 6.4. Spatial Matching Estimates over Different Tolerances



Finally, the matching summary statistics (Table 5.7, inset; Tables 7A/7B, Appendix A) show that the algorithm was able to generate reasonably good balance results, which effectively eliminates the hypothesis that the discontinuity arises from covariate imbalance.

6.4 What is the value of the PAV?

Having arrived at the end of the methodological tussle, I now am able to answer the question of whether parents actually pay more to be situated close to good schools. The estimates from the three tolerance levels are presented in Table 6.5 (inset). I tabulate the average estimate across each specification, as well as the number of significant estimates at the 5% significance level.

Table 6.5 Priority Admission Value Results

Data	Tolerance Level					
	0.75 SD		1 SD		1.25 SD	
	Estimate	*Significance*	*Estimate*	*Significance*	*Estimate*	*Significance*
Public Housing, 1km	-0.14%	1 / 10	0.72%	7 / 10	0.56%	3 / 10
Public Housing, 2km	1.17%	9 / 10	0.95%	8 / 10	0.88%	8 / 10
Private Housing, 1km	2.77%	9 / 10	4.40%	10 / 10	3.74%	8 / 10
Private Housing, 2km	1.76%	10 / 10	3.32%	9 / 10	1.97%	2 / 10

NOTES:

- Significance: Number of Significant estimates for the given tolerance level, across 10 bandwidths

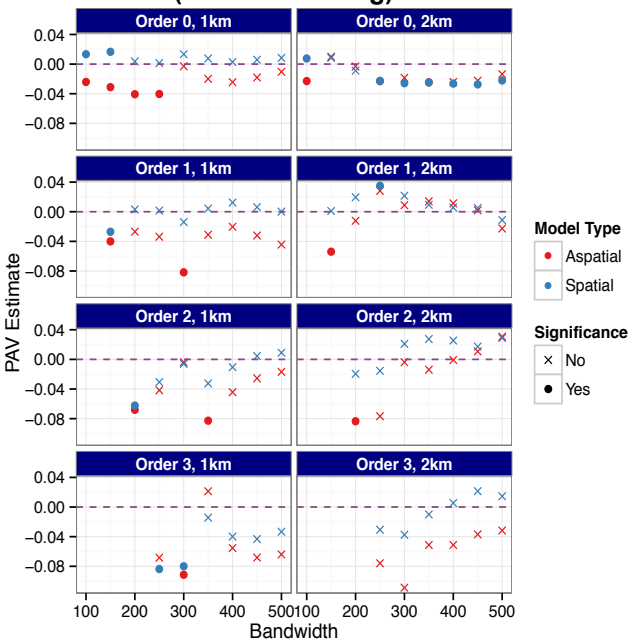
The results show a consistent positive estimate for private housing at the 1km boundary, with an overall average of 3.6%, or $58,240 if evaluated at median P1 price. The other consistently significant effect is public housing at the 2km boundary, with an overall average of 1.0%, or $4,450 if evaluated at median H2 price. The other somewhat significant effect is private housing at the 2km boundary, with an overall average of 2.35%, or $31,840, if evaluated at the median P2 price. There is no strong evidence of an effect at the H1 boundary.

7. Conclusions

7.1 Space and its relation to the models

On the whole, the discontinuity estimation produced surprising results. Although this alone does not provide a basis for rejecting the validity of its estimates, the myriad of choices available – bandwidth, order, kernel choice, residual type – produced differing results under each of the specifications, which is observed in Figures 6.2 and 6.5 (inset, below). Each of these four choices involves modeling spatial effects in some way. There may well be some ideal model that approximates the truth, but it is difficult to know how to find it, given that checking for consistency of results has failed along each of these dimensions.

Figure 6.5. Parametric RD Estimates (Private Housing)



More importantly, the persistent spatial trends of high spatial autocorrelation, non-monotonic evolution of land value, and the discontinuous density across the housing landscape induced by high-rise living, are all probably more than variations along these four dimensions can overcome. Up till now, the neighborhoods to which the RD approach has been applied to have been lower in population density, more homogeneous, and had more

cleanly defined boundaries, which may explain why previous papers have been successful in identifying the effect of interest.

Matching has proved to be a much more effective tool in side-stepping the thorny issue of space. Upon closer examination, the main pitfall of RD estimation lies in the non-comparability of the treatment and control groups, an issue that is not solved either by increasing sample size or bandwidth, and which is tackled straight on by matching. Unlike in the RD setting where the choices are on occasion rather arbitrary and hard to evaluate, there are more logically and theoretically sound ways to decide on the parameterization of the matching approach, such as looking at covariate balance, and assessing the appropriateness of the actual matching assignment.

7.2 The Present and the Future

Taking the matching results as the benchmark, many Singaporean parents are willing to pay significant sums of money for priority admission to schools with Gifted Education Programmes. If Wong's (2011) results are taken at face value, then the small to insignificant public housing estimates produced by this dataset are a significant drop from his. This might suggest, preliminarily, that the MOE's policy of encouraging equality amongst schools in Singapore, at least from a revealed preference perspective, has had a discernible effect.

This research warrants further extensions to the spatial analysis here, most of which has yet to exploit the full arsenal of spatial econometric tools developed in recent years. It would require a much more sophisticated geo-statistical model such as a spatial autoregressive moving average (SARMA) model, or at least the use of more developed diagnostic tools for model selection. Another area to look into is school-specific analysis in an environment where school boundaries overlap. With a smaller sample size, linear models will tend to experience difficulties in this regard. If matching approaches can circumvent this limitation, this would provide an alternate way to measure school valuation.

References

Aaronson, D., Barrow, L. & William, S. (2007). "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, 25(11), 95-135.

Acemoglu, D. & Angrist, J. (2000). How Large are Human-Capital Externalities? Evidence from Compulsory-Schooling Laws. In Bernanke, B. & Rogoff, K. (Eds.), *NBER Macroeconomics Annual 2000* (Vol. 15, pp. 9-74). Cambridge, MA: MIT Press.

Agarwal, S., Rengarajan, S., & Sing, T. F. (2014). *Values of Proximity to Schools: An Experiment with School Relocation Events in Singapore* (Working Paper). Social Science Research Network. Retrieved December 16, 2014, from: http://ssrn.com/abstract=2380761

Anselin, L. (2009). *Thirty Years of Spatial Econometrics* (Working Paper No. 2009-02). Phoenix, AZ: GeoDa Center for Geospatial Analysis and Computation. Retrieved December 16, 2014, from: https://geodacenter.asu.edu/ system/files/Anselin0902.pdf

Anselin, L. (2001). Spatial Econometrics. In Baltagi, B., (Ed.), *A Companion to Theoretical Econometrics* (pp. 310–330). Blackwell: Oxford.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Araujo, M., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2014). *A Helping Hand? Teacher Quality and Learning Outcomes in Kindergarten.* (Working Paper). Washington, DC: Inter-American Development Bank.

Bayer, P., Ferreira, F., & McMillan, R. (2007). "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy*, 115(4), 588-638.

Black, S. (1999). "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics*, 114(2), 577-599.

Black, S., Devereux, P., & Salvanges, K. (2005). "The More the Merrier? The Effect of Family Size on Birth Order on Children's Education." *The Quarterly Journal of Economics*, 120(2), 669-700.

Black, S. & Machin, S. (2010). Housing Valuations of School Performance. In Hanushek, E., Machin, S., & Woessmann, L. (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 486-516), Amsterdam: Elsevier.

Bogart, W. & Cromwell, B. (1997). "How Much More is a Good School District Worth?" *National Tax Joural*, 50(2), 215-232.

Calonico, S., Cattaneo, M., & Titiunik, R. (2013). "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." Unpublished Manuscript.

Carneiro, P., Meghir, C., & Parey, M. (2013). "Maternal Education, Home Environments, and the Development of Children and Adolescents." *Journal of the European Economic Association*, 11(S1), 123-160.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review*, 101(6): 2754-81.

Chia, S. (2013, July 17). Primary 1 registration: Few Places Left in Popular Schools. Retrieved December 16, 2014, from: http://www.straitstimes.com/the-big-story/ask-sandra-p1-registration/story/primary-1-registration-few-places-left-popular-school

Ciccone, A. & Peri, G. (2002). *Identifying Human Capital Externalitie: Theory with an Application to US Cities*. (Discussion Paper No. 488). Bonn: Institute for the Study of Labor. Retrieved December 16, 2014, from: http://ftp.iza.org/dp488.pdf

Cliff, A. & Ord, J. (1973). *Spatial Autocorrelation*. London: Pion.

Clotfelter, C., Ladd, H. & Vigdor, J. (2010). "Teacher Credentials and Student Achievement in High School." *Journal of Human Resources*, 45(3), 655-681.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Davie, S. (2014, August 7). Move Top Schools out of Bukit Timah. Retrieved December 16, 2014, from: http://www.straitstimes.com/news/opinion/eye-singapore/story/move-top-schools-out-bukit-timah-20140807

DesJardins, S. & McCall, B. (2008). "The Impact of the Gates Millenium Scholars Program on the Retention, College Finance- and Work-related Choices, and Future Educational Aspirations of Low-Income Minority Students." *Unpublished Manuscript*.

Ehrenberg, R. & Brewer, D. (1994). "Do School and Teacher Characteristics Matter? Evidence from High School and Beyond." *Economics of Education Review*, 13(1), 1-17.

Fack, G. & Grenet, J. (2010). "When do Better Schools Raise Housing Prices? Evidence from Paris Public and Private Schools." *Journal of Public Economics*, 94(1), 59-77.

Ferguson, R. & Brown, J. (2000). Certification Test Scores, Teacher Quality and Student Achievement. In Grissmer, D. & Ross, J. (Eds.), *Analytic Issues in the assessment of student achievement* (pp. 133-156). Washington, DC: National Center for Education Statistics.

Gibbons, S. & Machin, S. (2003). "Valuing English Primary Schools." *Journal of Urban Economics*, 53(2), 197-219.

Gibbons, S., Machin, S., 2008. Valuing school quality, better transport and lower crime: evidence from house prices. Oxford Review of Economic Policy 24, 99–119.

Gibbons, S., Machin, S. & Silva, O. (2013). "Valuing School Quality using Boundary Discontinuities." *Journal of Urban Economics*, 75(2), 15-28.

Gibbons, S., Machin, S. & Viarengo, M. (2012). *Does Additional Spending Help Urban Schools? An Evaluation Using Boundary Discontinuities.* (Discussion Paper No. 6281). Bonn: Institute for the Study of Labor. Retrieved December 16, 2014, from: http://ftp.iza.org/dp6281.pdf

Griliches, Z. (1990). Hedonic Price Indexes and the Measurement of Capital and Productivity: Some Historical Reflections. In Berndt, E. & Triplett, J. (Eds.), *Fifty Years of Economic Measurement: The Jubilee of the Conference on Research in Income and Wealth* (pp. 185-206). Chicago, IL: University of Chicago Press.

Hanushek, E. (1986). "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, 24(3), 1141-1177.

Hanushek, E. (2005). "The Economics of School Quality." *German Economic Review*, 6(3), 269-286.

Haurin, D. & Brasington, D. (1996). *The Impact of School Quality on Real House Prices: Interjurisdictional Effects*. (Working Paper No. 010). Athens, OH: Ohio State University. Retrieved November 20, 2014, from: http://ecolan.sbs.ohio-state.edu/pdf/haurin/ haurin.pdf

Heng, L. (2013, November 29). Wisdom of the masses or just plain kiasu? Retrieved April 28, 2015, from: http://news.asiaone.com/news/singapore/wisdom-masses-or-just-plain-kiasu

Imbens, G. & Kalyanaraman, K. (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies*, 79(3), 933-959.

Imbens, G. & Lemieux, T. (2008). "Regression Discontinuity Design: A Guide to Practice." *Journal of Econometrics*, 142(2), 615-635.

Kane, T., Taylor, E., Tyler, J. & Wooten, A. (2011). "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources*, 46(3), 587-613.

Keele, L. & Titiunik, R. (2015). Geographic Boundaries as Regression Discontinuities. *Political Analysis*, 23(1), 127-155.

Ladd, H. & Loeb, S. (2014). The challenges of measuring school quality: implications for educational equity. In Allen, D. & Reich, D. (Eds.), *Education, Justice and Democracy* (pp. 22-55). Chicago, IL: University of Chicago Press.

Lee, D. & Lemieux, T. (2010). "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2), 281-355.

Lee, P. (2014, July 22). Despite New Cooling Measure, More Primary Schools Oversubscribed at Phase 2B this Year. Retrieved December 16, 2014, from: http://www.straitstimes.com/news/singapore/education/story/despite-new-cooling-measure-more-primary-schools-oversubscribed-phase

Lleras-Muney, A. (2005). "The Relationship between Education and Adult Mortality in the United States." *Review of Economic Studies*, 72(1), 189-221.

Lochner, L. & Moretti, E. (2004). "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests and Self-Reports." *American Economic Review*, 94(1), 155-189.

Loeb, S. & Page, M. (2000). "Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor-Market Opportunities and Non-Pecuniary Variation." *The Review of Economics and Statistics*, 82(3), 393-408.

Ludwig, J. & Miller, D. (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics*, 122(1), 159-208.

Machin, S. (2011). "Houses and Schools: Valuation of School Quality Through the Housing Market." *Labour Economics*, 18(6), 723-729.

Meghir, C., Palme, M. & Simeonova, E. (2012). *Education, Health and Mortality: Evidence from a Social Experiment*. (Working Paper No. 17932). Cambridge, MA: National Bureau of Economic Research. Retrieved December 16, 2014, from: http://www.nber.org/papers/ w17932.pdf

Milligan, K., Moretti, E. & Oreopoulos, P. (2003). *Does Education Improve Citizenship? Evidence from the U.S. and the U.K.* (Working Paper No. 9584). Cambridge, MA: National Bureau of Economic Research. Retrieved December 16, 2014, from: https://ideas.repec.org/p/nbr/ nberwo/ 9584.html

Ministry of Education. (2013). Every School a Good School. Retrieved February 4, 2015, from: http://www.moe.gov.sg/initiatives/every-school-good-school/

Ministry of Education. (2014). Gifted Education Programme: Gifted Education Programme Schools. Retrieved April 28, 2015, from: http://www.moe.gov.sg/education/programmes/gifted-education-programme/gep-schools/

Moretti, E. (2004). "Estimating the Social Return to Highe Education: Evidence from Longitudinal and Repeated Cross-Sectional Data." *Journal of Econometrics*, 121(1-2), 175-212.

Prime Minister's Office. (2013). Prime Minister Lee Hsien Loong's National Day Rally 2013 (Speech in English). Retrieved February 4, 2015, from: http://www.pmo.gov.sg/mediacentre/prime-minister-lee-hsien-loongs-national-day-rally-2013-speech-english

Rivkin, S., Hanushek, E. & Kain, J. (2005). "Teachers, Schools and Academic Achievement." *Econometrica*, 73(2), 417-458.

Rosen, S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, 82(1), 34-55.

Rosenbaum, P. & Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1), 41-55.

Rubin, D. (1974). "Estimating Causal Effects of Treatment in Randomized and non-Randomized Studies." *Journal of Educational Psychology*, 66(5), 688-701.

Sekhon, J. (2011). "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal of Statistical Software*, 42(7).

Sheppard, S. (1999). Hedonic Analysis of Housing Markets. In Cheshire, P. & Mills, E. (Eds.), *Handbook of Regional and Urban Economics* (1st ed., Vol. 3, pp. 1595-1635), Elsevier.

The Straits Times. (2014). Take Up Rate at all Primary Schools in 2013. Retrieved April 8, 2015, from: http://www.straitstimes.com/STI/STIMEDIA/2014/school_infographics/index.html

Taylor, L. & Fowler, W. (2006). *A Comparable Wage Approach to Geographic Cost Adjustment* (Research and Development Report No. 321). Washington, DC: National Center for Education and Statistics. Retrieved December 16, 2014, from: http://nces.ed.gov/pubs2006/2006321.pdf

Tiebout, C. (1956). "A Pure Theory of Local Expenditures." *Journal of Political Economy*, 64(5), 416-424.

Thistlewaite, D. L. & Campbell, D. T. (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology*, 51(6), 309-317.

Wong, W. K. (2011). *Parental Valuation of Priority Admission to Primary Schools: The Effects of Academic Reputation and Choices.* (Working Paper). Singapore: National University of Singapore. Retrieved November 20, 2014, from http://courses.nus.edu.sg/course/ecswong/ workingpapers/wkwong87.pdf

Appendix A: Tables and Figures

Table 1. Observation Selection

Global Analysis	Dropped	Total Left
Private Housing Transactions		
Selected only transactions involving apartment and condominium type housing for the years 2012-2014		21,682
Kept only observations between March 2012 and November 2014	998	20,684
Dropped observations without geo-codes	49	20,635
Dropped observations without completion date	689	19,946
Dropped outlier observations with floor area more than 300 square meters	272	**19,674**
Public Housing Transactions		
Used all observations made available on data.gov.sg as of December 2014		50,936
Dropped 1-room and 2-room flats	606	**50,330**
Global Analysis: Total Housing Transactions		**70,004**

Discontinuity Analysis	Dropped	Total Left
Private Housing Transactions		
Selected observations within 500m of 1km Boundary		5,216
Dropped observations with no support across the boundary	150	5,066
Selected observations within 500m of 2km Boundary		8,992
Total Private Housing Transactions Used:		**14,058**
Public Housing Transactions		
Selected observations within 500m of 1km Boundary		3,807
Dropped observations with no support across the boundary	258	3,549
Selected observations within 500m of 2km Boundary		4,478
Total Public Housing Transactions Used:		**8,027**
Discontinuity Analysis: Total Housing Transactions Used		**22,085**

Table 2. Descriptive Statistics for Public and Private Housing Data

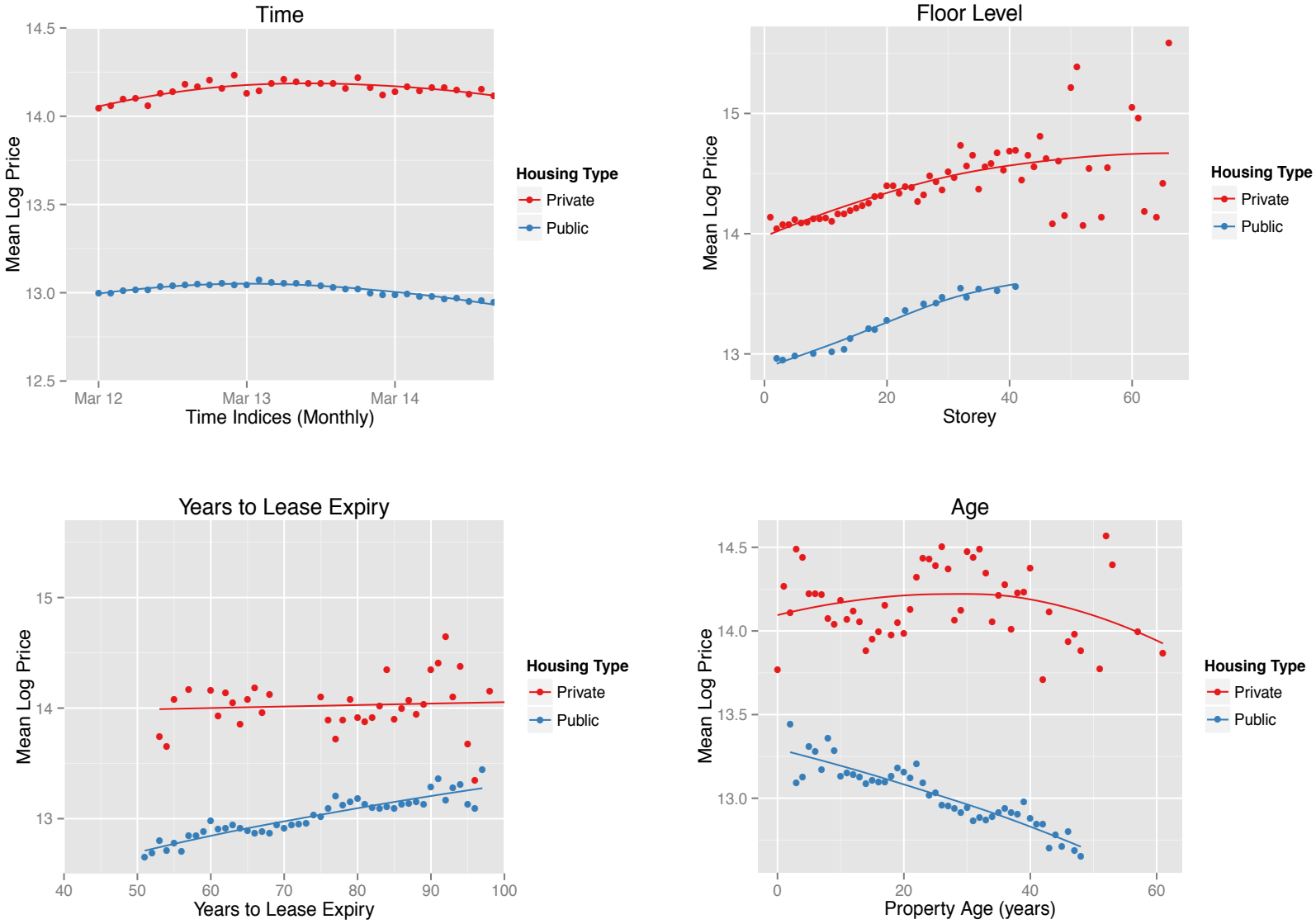
Observational Characteristics	Full Sample		Public		Private	
Summary Statistics	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Continuous Variables						
Price (in thousands of SGD)	770	713	464	121	1,550	962
Log Price	13.3	0.6	13.0	0.2	14.2	0.5
Floor Area (square meters)	104	32	97	24	121	41
Floor level	7.8	5.6	7.7	4.9	8.3	7.0
Age	21.1	10.8	24.3	10.0	12.8	7.9
Years to Lease Expiry	205	317	75	10	438	452
Distance to MRT	797	454	783	418	831	535
Number of Boundaries per Observation*	10.30	4.07	10.87	3.98	8.88	3.93
Summary Statistics			**Count**	%	**Count**	%
Flat Type*						
Apartment			3,505	6.96%		
Improved			13,154	26.14%		
Model A			14,904	29.61%		
New Generation			9,354	18.59%		
Premium Apartment			4,595	9.13%		
Simplified			3,017	5.99%		
Standard			1,801	3.58%		
Freehold Apartment					3,628	18.44%
Leasehold Apartment					2,032	10.33%
Freehold Condominium					5,105	25.95%
Leasehold Condominium					7,422	37.72%
Leasehold Executive Condominium					1,487	7.56%

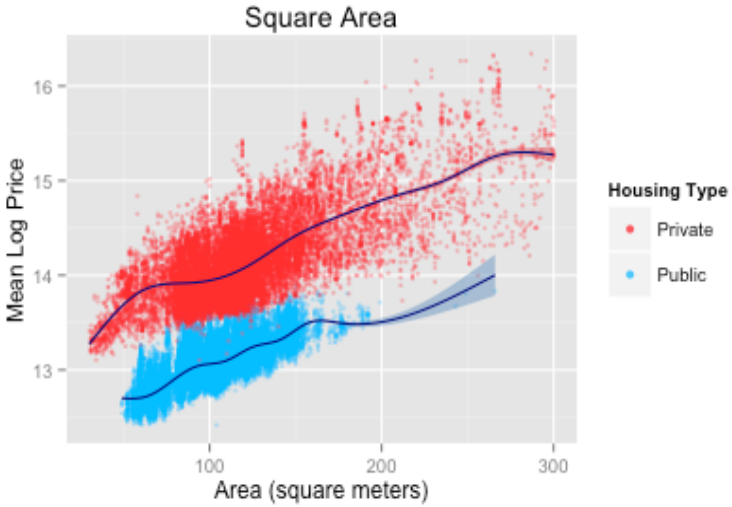
NOTES:

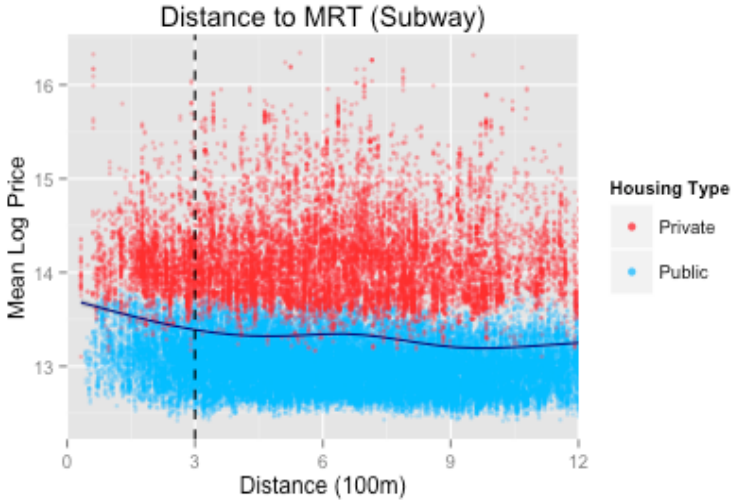
*** The first 7 categories are for Public Housing and the next 5 for Private Housing**

Figure 3. Summary Statistics

Plots of Covariates against Price

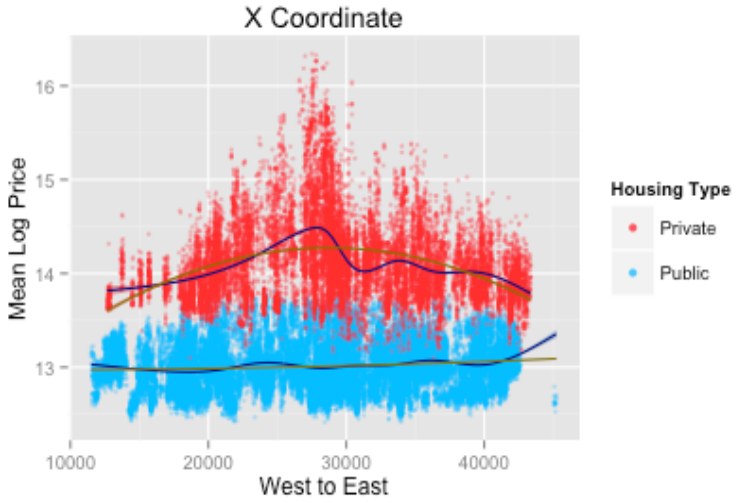


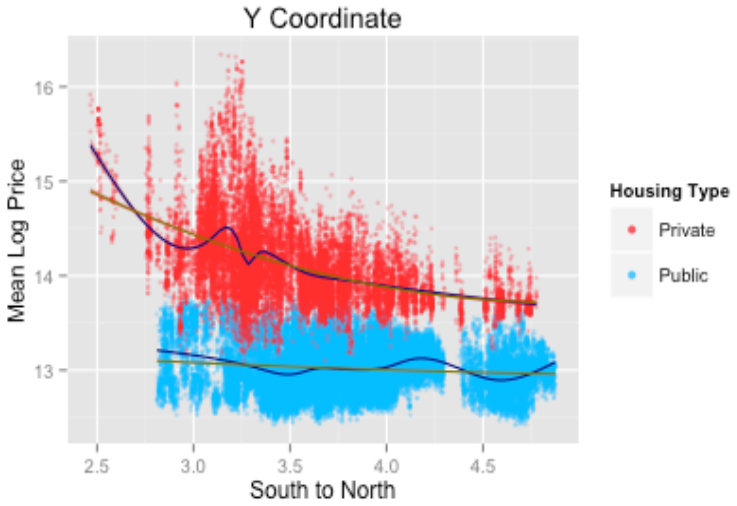

Square Area
Mean Log Price
16
15
14
13
100
200
300
Area (square meters)
Housing Type
Private
Public


Distance to MRT (Subway)
Mean Log Price
16
15
14
13
0
3
6
9
12
Distance (100m)
Housing Type
Private
Public

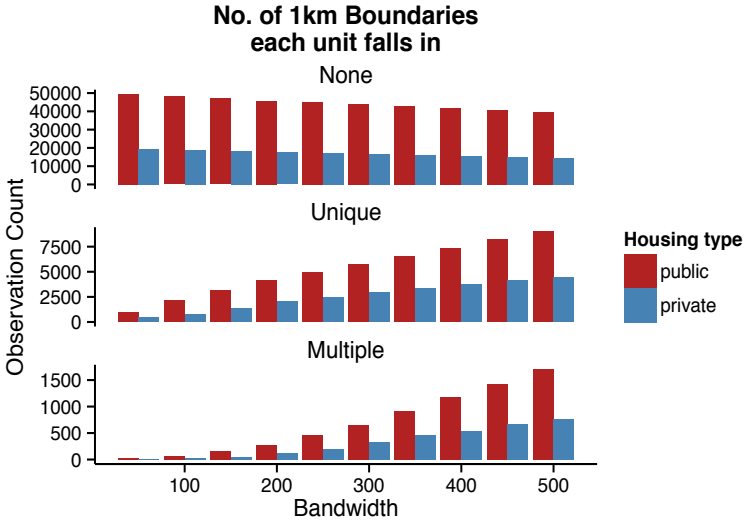

Housing Type
Mean Price ($)
2,000,000
1,500,000
1,000,000
500,000
0
Housing Type
Housing Type
Apartment
Premium Apartment
Improved
Model A
Standard
Simplified
New Generation
Freehold Condo
Freehold Apartment
Leashold Condo
Leasehold Apartment
Leasehold Exec Condo

Plots of Geo-Coordinates Against Price





Numbers of Observations within 500m of 1km and 2km Boundaries



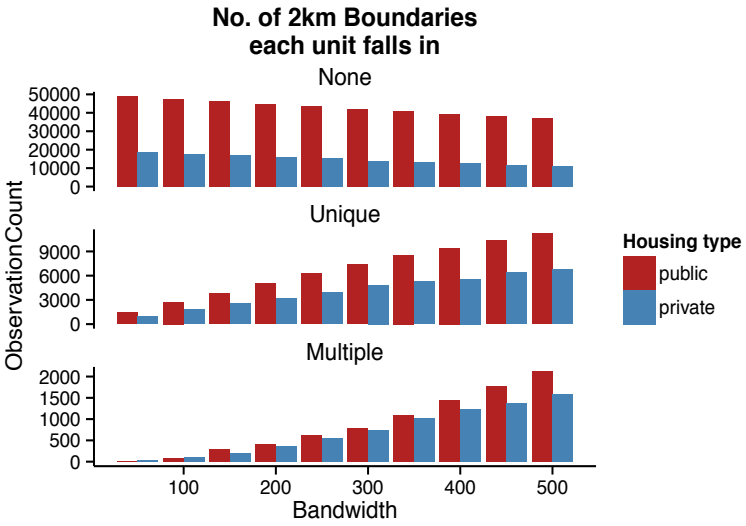


Table 4A. HDB Regression Coefficients on Log Housing Price

Regression Models	**HDB (1)**	**HDB (2)**	**HDB (3)**	**HDB (4)**	**HDB (5)**
N = 19,674	**Baseline Specification**	**Add Y Coord trend (deg 1)**	**Add X Coord trend (deg 1-2)**	**Add Y Coord trend (deg 2)**	**Add X Coord trend (deg 3-4)**
Intercept (Flat type: HDB Apartment)	12.325***	12.991***	12.450***	12.942***	13.628***
Square Area	0.009***	0.009***	0.009***	0.009***	0.009***
Floor Level	0.010***	0.009***	0.008***	0.008***	0.008***
Age	-0.030***	-0.022***	-0.015***	-0.015***	-0.016***
Age (degree 2)	0.001***	0.0004***	0.0002***	0.0002***	0.0002***
Time	0.011***	0.010***	0.010***	0.010***	0.010***
Time (degree 2)	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***
Near to MRT Station (Binary)	0.169***	0.086***	0.078***	0.071***	0.064***
Flat type: HDB Premium Apartment	-0.031***	0.004	0.018***	0.020***	0.022***
Flat type: HDB Improved	0.008***	0.007***	-0.005	-0.002	-0.002
Flat type: HDB Model A	0.030***	0.021***	0.018***	0.018***	0.015***
Flat type: HDB Standard	-0.019***	-0.013***	-0.023***	-0.027***	-0.025***
Flat type: HDB Simplified	0.077***	0.095***	0.057***	0.058***	0.056***
Flat type: HDB New Generation	0.057***	0.062***	0.044***	0.045***	0.046***
X Coordinate			0.489***	0.459***	-1.451***
Y Coordinate		-0.175***	-0.220***	-0.464***	-0.264***
X Coordinate (Degree 2)			-0.081***	-0.076***	1.150***
Y Coordinate (Degree 2)				0.031***	0.005**
X Coordinate (Degree 3)					-0.328***
Y Coordinate (Degree 3)					
X Coordinate (Degree 4)					0.031***
Y Coordinate (Degree 4)					
Adjusted R Squared	**0.74**	**0.82**	**0.87**	**0.87**	**0.88**
Residual Sum of Squares	**814.48**	**551.24**	**391.97**	**389.47**	**381.58**
F-Statistic		**34,705*** **	**10,499*** **	**329*** **	**520*** **

Note: * significant at 10%, ** significant at 5%, * significant at 1%**

Table 4B. PP Regression Coefficients on Log Housing Price

Regression Models	**PP (1)**	**PP (2)**	**PP (3)**	**PP (4)**	**PP (5)**
N = 50,330	**Baseline Specification**	**Add Y Coord trend (deg 1-2)**	**Add X Coord trend (deg 1-2)**	**Add X Coord trend (deg 3-4)**	**Add Y Coord trend (deg 3-4)**
Intercept (Flat type: Freehold Apartment)	13.116***	16.739***	13.763***	18.338***	23.475***
Square Area	0.008***	0.007***	0.007***	0.007***	0.007***
Floor Level	0.011***	0.008***	0.006***	0.006***	0.006***
Age	-0.025***	-0.016***	-0.016***	-0.016***	-0.016***
Age (degree 2)	0.0004***	0.0002***	0.0002***	0.0002***	0.0002***
Time	0.017***	0.017***	0.017***	0.016***	0.016***
Time (degree 2)	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
Near to MRT Station (Binary)	0.049***	0.052***	0.037***	0.038***	0.039***
Years to Lease Expiry	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***
Flat type: Freehold Condominium	0.056***	0.112***	0.141***	0.127***	0.127***
Flat type: Leasehold Apartment	-0.034***	0.062***	0.062***	0.051***	0.054***
Flat type: Leasehold Condominium	0.029***	0.129***	0.138***	0.128***	0.129***
Flat type: Leasehold Executive Condominium	-0.248***	0.046***	0.090***	0.080***	0.082***
X Coordinate			0.762***	-7.883***	-7.843***
Y Coordinate		-1.721***	-0.601***	-0.288***	-6.392**
X Coordinate (Degree 2)			-0.137***	5.061***	5.040***
Y Coordinate (Degree 2)		0.185***	0.033***	-0.10*	2.655**
X Coordinate (Degree 3)				-1.322***	-1.317***
Y Coordinate (Degree 3)					-0.510**
X Coordinate (Degree 4)				0.121***	-0.120***
Y Coordinate (Degree 4)					0.036**
Adjusted R Squared	**0.70**	**0.80**	**0.83**	**0.84**	**0.84**
Residual Sum of Squares	**1,111.80**	**730.89**	**629.11**	**582.69**	**582.43**
F-Statistic		**6,426*** **	**1,717*** **	**783.2*** **	**4.41** **

Note: * significant at 10%, ** significant at 5%, * significant at 1%**

Table 4C. HDB Spatial Regression Coefficients on Log Housing Price

Model No.:	HDB (1)	HDB (2)	HDB (3)	HDB (4)	HDB (5)	HDB (6)	HDB (7)
(Housing Attributes omitted from this table)	Baseline	IDW 150 Lag	IDW 200 Lag	Binary 150 Lag	KNN (10)	KNN (12)	KNN (15)
X Coordinate	0.489***	0.379***	0.377***	0.387***	0.308***	0.314***	0.322***
X Coordinate (Degree 2)	-0.082***	-0.064***	-0.064***	-0.065***	-0.052***	-0.053***	-0.054***
Y Coordinate	-0.220***	-0.161***	-0.160***	-0.165***	-0.130***	-0.132***	-0.132***
IDW 150m Lag		-0.299***					
IDW 200m Lag			0.319***				
Binary 150m Lag				0.288***			
KNN-10 Lag					0.429***		
KNN-12 Lag						0.422***	
KNN-15 Lag							0.410***
Adjusted R Squared	**0.87**	**0.90**	**0.90**	**0.89**	**0.92**	**0.91**	**0.91**
Average Moran's I	**0.56**	**0.37**	**0.37**	**0.38**	**0.32**	**0.32**	**0.33**
Residual Sum of Squares (over Baseline)	**393.32**	**69.77**	**69.57**	**64.99**	**125.49**	**120.92**	**112.54**
F-Statistic (compared to Baseline)		**10,848**	**10,811**	**9,958**	**23,572**	**22,333**	**20,166**

Note: * significant at 10%, ** significant at 5%, * significant at 1%**

Table 4D. Spatial PP Regression Coefficients on Log Housing Price

Model No.:	PP (1)	PP (2)	PP (3)	PP (4)	PP (5)	PP (6)	PP (7)
(Housing attributes omitted from this table)	Baseline	IDW 150 Lag	IDW 200 Lag	Binary 200 Lag	KNN-10 Lag	KNN-12 Lag	KNN-15 Lag
X Coordinate	0.762***	0.417***	0.404***	0.413***	0.295***	0.295***	0.294***
X Coordinate (Degree 2)	-0.137***	-0.075***	-0.072***	-0.074***	-0.053***	-0.053***	-0.052***
Y Coordinate	-0.601***	-0.131***	-0.114***	-0.105***	-0.215***	-0.194***	-0.196***
Y Coordinate (Degree 2)	0.033***	-0.008	-0.009*	-0.011**	0.009**	0.007	0.007
IDW 150m Lag		0.398***					
IDW 200m Lag			0.414***				
Binary 200m Lag				0.433***			
KNN-10 Lag					0.554***		
KNN-12 Lag						0.557***	
KNN-15 Lag							0.558***
Adjusted R Squared	**0.87**	**0.89**	**0.90**	**0.89**	**0.92**	**0.92**	**0.92**
Average Moran's I Statistic	**0.53**	**0.17**	**0.16**	**0.18**	**0.22**	**0.22**	**0.21**
Residual Sum of Squares (over Baseline)	**629.14**	**231.41**	**236.42**	**228.83**	**319.75**	**320.59**	**319.36**
F-Statistic (compared to Baseline)		**11,436**	**11,833**	**11,236**	**20,315**	**20,423**	**20,263**

Note: * significant at 10%, ** significant at 5%, * significant at 1%**

Figure 5. Plots of binned covariate means (over 50m bins)

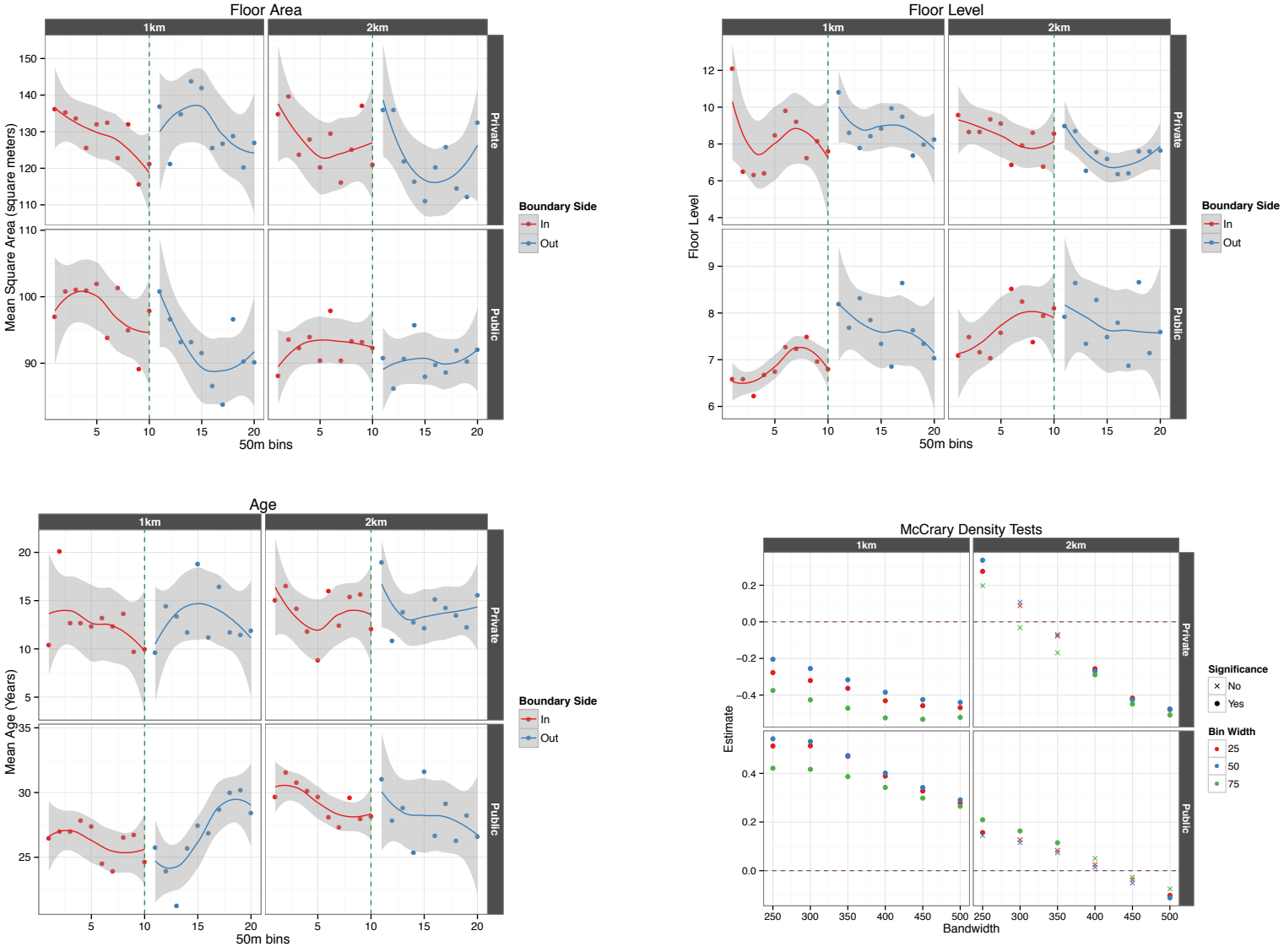


Figure 6. Polynomial Fits for the four datasets

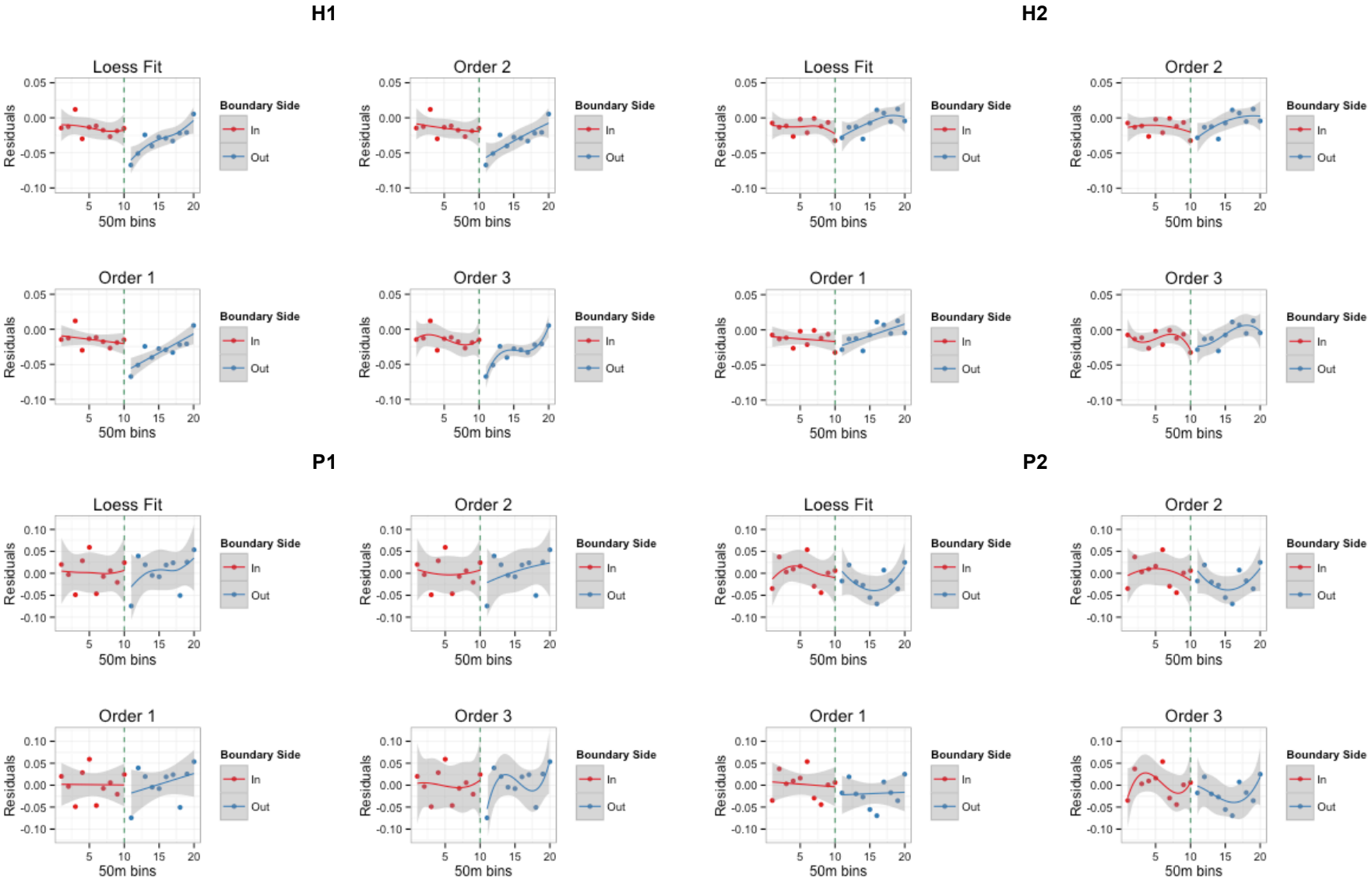


Table 7A. Matching Balance Statistics (Public Housing)

	Matching Type	Unmatched			Coordinates			Covariates			Covariates & Coordinates		
Data	**Variable**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**
H1	Age	0.74	0.00	0.00	0.89	0.00	0.00	1.26	0.00	0.00	1.19	0.00	0.00
	Age (Squared)	0.75	0.00	0.00	1.07	0.00	0.00	1.16	0.00	0.00	1.11	0.00	0.00
	Square Area	1.26	0.00	0.00	0.95	0.00	0.00	1.02	0.00	0.02	1.01	0.20	0.06
	Floor Level	0.60	0.00	0.00	0.43	0.00	0.00	1.03	0.00	0.93	1.05	0.00	0.51
	Time	1.02	0.04	0.08	1.04	0.00	0.00	1.00	0.09	0.99	1.05	0.05	0.60
	Time (Squared)	1.04	0.03	0.08	1.06	0.00	0.00	1.01	0.13	0.99	1.06	0.00	0.60
	MRT Proximity (Binary)	1.15	0.17		0.64	0.00		1.01	0.32		1.03	0.03	
	Type (Apartment)	2.88	0.00		1.19	0.04		1.00	1.00		1.00	1.00	
	Type (Improved)	1.00	0.90		1.04	0.34		1.00	1.00		0.98	0.00	
	Type (Model A)	1.02	0.33		0.96	0.01		1.00	0.05		1.00	1.00	
	Type (New Generation)	1.03	0.45		1.09	0.02		1.00	1.00		1.03	0.00	
	Type (Premium Apartment)	1.19	0.26		141.16	0.00		1.00	1.00		1.00	1.00	
	Type (Simplified)	0.50	0.00		0.73	0.00		1.03	0.05		1.00	1.00	
	Type (Standard)	0.11	0.00		0.11	0.00		1.00	1.00		1.00	1.00	
	Log Price	0.85	0.01	0.00	0.89	0.00	0.00	0.93	0.93	0.00	1.17	0.00	0.00
H2	Age	0.71	0.00	0.00	0.60	0.00	0.00	1.08	0.00	0.00	1.09	0.00	0.00
	Age (Squared)	0.82	0.00	0.00	0.68	0.00	0.00	1.16	0.00	0.00	1.14	0.00	0.00
	Square Area	0.87	0.00	0.00	0.80	0.00	0.00	1.00	0.02	0.11	1.03	0.00	0.07
	Floor Level	0.89	0.37	0.57	1.18	0.00	0.00	1.10	0.00	0.34	1.17	0.00	0.11
	Time	1.01	0.09	0.07	0.99	0.00	0.00	1.00	0.73	0.86	1.03	0.27	0.30
	Time (Squared)	1.00	0.15	0.07	0.94	0.00	0.00	1.00	0.97	0.86	1.03	0.01	0.30
	MRT Proximity (Binary)	1.05	0.35		0.88	0.00		1.00	1.00		1.00	1.00	
	Type (Apartment)	0.85	0.06		0.49	0.00		1.00	1.00		1.00	1.00	
	Type (Improved)	0.80	0.00		0.98	0.31		1.00	1.00		1.00	1.00	
	Type (Model A)	1.07	0.03		0.80	0.00		1.00	1.00		1.00	1.00	
	Type (New Generation)	1.09	0.00		1.18	0.00		1.00	1.00		1.00	1.00	
	Type (Premium Apartment)	1.50	0.00		22.76	0.00		1.00	1.00		1.00	1.00	
	Type (Simplified)	1.07	0.45		0.67	0.00		1.00	1.00		1.00	1.00	
	Type (Standard)	0.78	0.04		1.18	0.26		1.00	1.00		1.00	1.00	
	Log Price	0.75	0.00	0.00	0.76	0.00	0.00	0.89	0.36	0.00	0.99	0.14	0.22

NOTES: The three statistics presented are the variance ratio as a measure of equality in variances, the T-test for equality of means, and the Kolmogorov-Smirnov (KS) test for equality of distributions. For the two tests, higher P-values indicate better balance. For the variance ratio, measures close to 1 indicate better balance. The KS Test is not meaningful for binary variables.

Table 7B. Matching Balance Statistics (Private Housing)

	Matching Type	Unmatched			Coordinates			Covariates			Covariates & Coordinates		
Data	**Variable**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**	**Var Ratio**	**T-Test**	**KS-Test**
P1	Age	0.78	0.19	0.00	0.67	0.00	0.00	1.21	0.19	0.00	1.20	0.80	0.00
	Age (Squared)	0.69	0.01	0.00	0.49	0.00	0.00	1.36	0.00	0.00	1.35	0.00	0.00
	Square Area	0.84	0.16	0.01	0.88	0.00	0.00	1.20	0.65	0.09	1.24	0.56	0.00
	Floor Level	0.95	0.01	0.01	1.26	0.00	0.00	1.15	0.00	0.41	1.24	0.00	0.00
	Time	1.00	0.73	0.80	0.93	0.04	0.00	1.00	0.00	0.10	1.00	0.00	0.00
	Time (Squared)	0.99	0.80	0.80	0.89	0.03	0.00	0.99	0.01	0.10	1.00	0.00	0.00
	MRT Proximity (Binary)	1.04	0.76		1.68	0.00		1.00	1.00		1.00	1.00	
	Type (Freehold Apartment)	0.85	0.00		0.82	0.00		1.00	1.00		1.00	1.00	
	Type (Freehold Condo)	1.08	0.00		1.11	0.00		1.00	1.00		1.00	1.00	
	Type (Leasehold Apartment)	0.04	0.00		0.04	0.00		1.00	1.00		1.00	1.00	
	Type (Leasehold Condo)	1.55	0.00		2.15	0.00		1.00	1.00		1.00	1.00	
	Type (Leasehold Exec Condo	0.00	0.00		0.00	0.00		NA	1.00		NA	1.00	
	Log Price	0.55	0.00	0.00	0.92	0.54	0.00	0.86	0.10	0.00	0.88	0.00	0.00
P2	Age	1.15	0.51	0.00	0.77	0.00	0.00	1.11	0.00	0.00	1.28	0.00	0.00
	Age (Squared)	1.12	0.52	0.00	0.53	0.00	0.00	1.18	0.00	0.00	1.34	0.00	0.00
	Square Area	1.15	0.00	0.00	1.17	0.08	0.00	1.17	0.00	0.00	1.23	0.00	0.00
	Floor Level	1.33	0.00	0.00	1.68	0.00	0.00	1.21	0.00	0.16	1.29	0.00	0.01
	Time	1.06	0.32	0.09	1.06	0.19	0.00	1.04	0.54	0.14	1.05	0.86	0.04
	Time (Squared)	1.04	0.73	0.09	0.95	0.57	0.00	1.01	0.04	0.14	1.03	0.08	0.04
	MRT Proximity (Binary)	0.77	0.00		1.41	0.00		1.04	0.00		1.05	0.00	
	Type (Freehold Apartment)	1.27	0.00		0.86	0.00		1.00	1.00		1.00	0.32	
	Type (Freehold Condo)	1.17	0.00		1.13	0.00		1.00	1.00		1.00	0.32	
	Type (Leasehold Apartment)	0.57	0.00		1.13	0.27		1.00	1.00		1.01	0.32	
	Type (Leasehold Condo)	0.92	0.00		1.23	0.00		1.00	1.00		1.00	0.32	
	Type (Leasehold Exec Condo	0.00	0.00		NA	1.00		NA	1.00		NA	1.00	
	Log Price	1.08	0.00	0.00	1.14	0.00	0.00	1.08	0.00	0.00	1.14	0.07	0.00

NOTES: The three statistics presented are the variance ratio as a measure of equality in variances, the T-test for equality of means, and the Kolmogorov-Smirnov (KS) test for equality of distributions. For the two tests, higher P-values indicate better balance. For the variance ratio, measures close to 1 indicate better balance. The KS Test is not meaningful for binary variables.

Appendix B1: Terminology relevant to Singapore

Balloting: A process of drawing lots or a lottery to decide the allocation of registration places in a given primary school, where the number of applicants exceeds the number of vacancies.

Housing Development Board (HDB): A statutory board of the Singaporean government that plans and develops public housing towns in Singapore. Apartment units provided through the HDB are referred to as *HDB flats*.

Ministry of Education (MOE): The department in the Singaporean government responsible for the "formulation and implementation of education policies."

Primary One Registration Exercise: Conducted annually, this exercise allocates every child to a primacy school for his or her primary education in the upcoming school calendar year.

Primary schools (known as *elementary schools* in some countries): Schools in which children receive primary education, typically beginning at age seven (*Primary One*) and lasting six years until age 12 (*Primary Six*).

Primary School Leaving Examination (*PSLE*): An examination administered by the MOE nation-wide to all students at the end of Primary Six

Private housing or *private property*: A term used to denote housing built and sold by private developers in Singapore. This would include all properties designated for residential purposes not built by the HDB, from apartment to terraced housing.

Public housing: A blanket term used to refer to all housing provided by the Singaporean government through the HDB. The HDB plans, develops and markets housing in so-called HDB towns, and is also responsible for the provision of neighborhood facilities and amenities for residents.

Appendix B2: Registration Phases in the Primary One Registration Exercise

School Registration Phases

Phase 1

- For a child with a sibling studying in the primary school of choice

Phase 2A(1)

- For a child whose parent is a former student of the primary school and who has joined the alumni association as a member
- For a child whose parent is a member of the School Advisory / Management Committee

Phase 2A(2)

- For a child whose parent or sibling has studied in the primary school of choice
- For a child whose parent is a staff member of the primary school of choice

Phase 2B

- For a child whose parent has joined the primary school as a parent volunteer and given at least 40 hours of voluntary service to the school within the last year
- For a child whose parent is a member endorsed by the church/clan directly connected with the primary school
- For a child whose parent is endorsed as an active community leader

Phase 2C

- For all children who are eligible for Primary One in the following year and not yet registered in a primary school

Phase 2C Supplementary

- For a child who is not yet registered in a primary school after Phase 2C

Phase 3

- For a child who still has not been registered from earlier phases
- For a child who is neither a Singapore Citizen or a Singapore Permanent Resident

Priority Rules

From Phases 2A to 2C Supplementary, the following priority rules apply whenever the number of applicants exceeds the number of vacancies:

- Singapore Citizens (SCs) have absolute priority over Singapore Permanent Residents (PRs) when balloting is necessary in a specific phase. SCs will be admitted ahead of PRs before home-school distance is considered.
- If, within each subcategory of applicants (SCs or PRs) there is an excess of applicants, then applicants living within 1km of the school get absolute priority, followed by those living within 2km of the school, followed by applicants from elsewhere.

Appendix B3: A measure of school reputation

Below is a list of the top websites and forums resulting from the first page of a Google search of "Singapore top primary schools". Only the original websites are cited, and blogs appropriating material originating from other websites are not included. All websites were retrieved as of December 16, 2014.

1. The Asian Parent (10 schools)
Year released: 2013
Criteria: Past performance, reader feedback
Web: http://sg.theasianparent.com/top-primary-schools-in-singapore-2013/

2. Google's Singapore Top Primary Schools (22 schools)
Year released: 2014
Criteria: "Pupils' performances in examinations"
Web: https://www.google.com/maps/d/u/0/viewer?oe=UTF8&t=h&ie=UTF8&msa=0&mid=zwFX1shloEnI.kFtK1mpoRb0g

3. SG Teach (21 schools)
Year released: 2014
Criteria: MOE Masterplan of Awards
Web: https://www.facebook.com/sgteach

4. Kiasu Parents: Academic Excellence (updated by GreatMinds) (20 schools)
Year released: 2011
Criteria: Students scoring above 275, top 10 national graduates
Web:
http://www.greatminds.edu.sg/index.php?option=com_content&view=article&id=85:singapore-primary-schools-ranked-by-academic-excellence&catid=59:great-minds-club&Itemid=91

5. Kiasu Parents: Popularity (20 schools)
Year released: 2008
Criteria: Number of applicants, take up rates

Web: http://www.kiasuparents.com/kiasu/content/singapore-primary-schools-ranked-popularity

6. Singapore Learner (21 schools)
Year released: 2013
Criteria: MOE Awards, number of Gifted Education Program classes
Web: http://singaporelearner.com/2013/01/08/list-of-top-primary-schools-2013-based-on-gep-classes-and-awards-achieved/

7. Edupoll (9 schools)
Year released: 2012
Criteria: Prestige
Web: http://www.edupoll.org/content/view/18/37/

Appendix B4: Computing school reputation scores

For each of the 7 websites, I assign the i^{th} school a score S_i which is positive if the school ranks in the top 20, and increasing with rank. In cases where there are no rankings I assume a tie between all schools, and where there are ties across the 20th rank I include all schools. The score is given by:

$$S_i = 21 - rank$$

So for instance, a school would net a score of 20 for ranking 1st on any of the websites. A list of the top 20 schools by score is presented below:

School (by score)	**Total Score**
Nanyang Primary School	112.5
Raffles Girls' Primary School	99.5
St. Hilda's Primary School	95
Rosyth School	89.5
Tao Nan School	84.5
Henry Park Primary School	83
Anglo-Chinese School (Primary)	81.5
Catholic High School	78.5
Rulang Primary School	64
Nan Hua Primary School	60
Singapore Chinese Girls' Primary School	42
Ai Tong School	40
Kong Hwa School	37
Methodist Girls' School (Primary)	36.5
CHIJ St. Nicholas Girls' School	33
Chongfu School	29
Nan Chiau Primary School	26.5
Pasir Ris Primary School	22.5
Tampines Primary School	21.5
Paya Lebar Methodist Girls' School (Primary)	21.5