NBER WORKING PAPER SERIES

ESTIMATING THE GAINS FROM NEW RAIL TRANSIT INVESTMENT: A MACHINE LEARNING TREE APPROACH

Seungwoo Chin
Matthew E. Kahn
Hyungsik Roger Moon

Working Paper 23326
http://www.nber.org/papers/w23326

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2017

Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach
Seungwoo Chin, Matthew E. Kahn, and Hyungsik Roger Moon
NBER Working Paper No. 23326
April 2017
JEL No. R21,R4

ABSTRACT

Urban rail transit investments are expensive and irreversible. Since people differ with respect to their demand for trips, their value of time, and the types of real estate they live in, such projects are likely to offer heterogeneous benefits to residents of a city. Using the opening of a major new subway in Seoul, we contrast hedonic estimates based on multivariate hedonic methods with a machine learning approach that allows us to estimate these heterogeneous effects. While a majority of the "treated" apartment types appreciate in value, other types decline in value. We explore potential mechanisms. We also cross-validate our estimates by studying what types of new housing units developers build in the treated areas close to the new train lines.

Seungwoo Chin
USC
Department of Economics
Los Angeles,CA 90089
chinseun@usc.edu

Hyungsik Roger Moon
USC
Department of Economics
Los Angeles,CA 90089
moonr@usc.edu

Matthew E. Kahn
Department of Economics
University of Southern California
KAP
Los Angeles, CA 90089
and NBER
kahnme@usc.edu

1 Introduction

Major urban rail transportation projects increase travel speeds in cities and thus facilitate shorter commutes, labor market matching and consumer shopping and leisure opportunities. Land close to new transport nodes often increases in value as the demand to live close to fast public transit increases local demand to live there. Real estate developers will seek to build new housing units close to these new stations.

In recent decades, Asia's major cities have made major investments in new subways (Gonzalez-Navarro and Turner (2016)). Cities ranging from Beijing, to Shanghai to Singapore have invested billions in subways. In this paper, we study how the real estate market in Seoul has been affected by the construction of a major new subway. The line number 9 (hereafter LINE9) subway connects the Southern part of the city with the Gangnam District. This is one of the richest parts of the city.

Our methodological approach builds on past hedonic studies that use panel estimation strategies to recover estimates of the causal effects of new transit access (Kahn (2007), Billings (2011), Zheng and Kahn (2013) and Gibbons and Machin (2005)). A distinguishing feature of our study is to use machine learning to pare down the possible non-linearities in the hedonic pricing function. Consider a hedonic regression that includes eleven explanatory variables that each takes on at least two discrete values. For example, one of such variables could be the apartment's size or an indicator for whether the apartment unit is located close to a new transit station. A researcher who seeks to flexibly estimate such a hedonic pricing gradient would need to include more than 2^{11} interaction terms. This is clearly infeasible but if the researcher does not pursue this strategy then the underlying pricing function may be misspecified.

Our solution to this challenge is to use the regression tree approach from machine learning (ML) (e.g., Breiman et al. (1984) and Friedman et al. (2001)). Building on Athey

and Imbens (2015), we apply ML methods in a difference-in-difference setting to estimate conditional average treatment effects. This approach imposes only light computational burdens. In our tree approach, we create dummy variables indicating whether the treatment has occurred or not and whether the housing unit is in the treatment area (i.e close to the new transit stations). The ML algorithm splits the sample on these attributes as well as on the physical attributes of the housing unit. This approach allows us to test how housing price appreciation differs for treated units versus control units while allowing these effects to vary by housing unit and community attributes. Earlier ML research has focused on predicting outcome variables using high-dimensional explanatory variables. A more recent literature has sought to use ML methods to estimate causal effects. These studies include, Zeileis et al. (2008), Beygelzimer and Langford (2009), Su et al. (2009), Foster et al. (2010), Dudík et al. (2011), Imai et al. (2013), Athey and Imbens (2015), and Taddy et al. (2016).

Based on our ML approach, we document that there is considerable variation in the conditional average treatment effect (the CATE). Some types of apartments experience greater price appreciation. For example, one "winner" from the treatment is an apartment in the upper 25% of the apartment size distribution featuring 3 rooms, 2 baths that is less than five years old and is located within one kilometer of old transit in the Seocho county. We contrast our ML estimates with the linear regression approach featuring a triple interaction term between a dummy for whether the treatment has taken place, and another dummy indicating whether the apartment is located in the treatment area. We then interact this paif or dummies with indicators for the apartment's physical attributes.

As a validation test of our estimates, we study whether developers of new apartments are building units with the features that our ML estimates predict yield the highest marginal revenue. We document a positive correlation between our estimates of the real estate price appreciation gains from train network proximity and the specific type of new housing

built by a developer. These findings support our claim that we have recovered key non-linearities of the true underlying pricing gradient and how they change over time.

2 Background

2.1 The New Subway Construction and Financing

Seoul's first subway line was built in 1974. Over the last four decades, the subway system expanded to cover five lines. In a continuing effort to mitigate congestion and to reduce commuting time, Seoul's government has built three additional subway lines since 2000 (the line number 6, 7 and 8). This expansion of the subway lines has contributed to an increase in the subway and rail utilization rates to 34.6% and 36.2% in 2002 and in 2010, respectively. The last subway expansion plan is the introduction of the LINE9. LINE9 was first designed in 1997. The detailed blueprint was released in 2000, and the ground-breaking construction ceremony took place in 2002. It began its service on July 24th in 2009. As of 2014, 39% of trips in Seoul use subways or railways. Baum-Snow and Kahn (2000) study the effects of sixteen different U.S cities' investments in new rail transit. The largest ridership gains are achieved when the new train is fast and connects to a city center where people want to go for work or consumption opportunities. Figure 1 and 2 display the network of earlier lines and LINE9. Figure 3 provides a chronology of the construction of this subway. A key assumption in a difference-in-difference approach is identifying the treatment date. The LINE9 was announced in the year 2000 but it was only completed years later. In section 5.1, we test and reject the hypothesis that the subway construction plant had an ex-ante capitalization effect.

The total cost of this project is US$818 millions[1]. 46.7% percent of the total costs are

[1]The construction costs are based on an exchange rate of 1100won/ US$1.

subsidized by the Seoul metropolitan government, and the METRO9, a private company, covered the rest. The METRO9 will operate the number 9 for three decades without paying any rental fees while the Seoul metropolitan government owns it. The Seoul metropolitan government guarantees a minimum profit levels for the first fifteen years of the project.[2] In 2005, the Korea Transport Institute predicted that 243,196 riders per a day in 2014 would use LINE9. However, the actual ridership has been 384,423 riders a day. This prediction stands in contrast to the U.S literature that argues that transit agencies routinely over-state the ridership of a new subway before it is built (Kain (1990)). Such strategic predictions increase the likelihood that the project is funded.

2.2 The Demand for Housing Close to Transit

Seoul's residents rely on public transit. In 2014, cars accounted for 22.8% and buses accounted for 27%, while subway and light rail take 39%. The share of trips by taxi is 6.8%.

Standard network logic suggests that the value of subway access increases in the set of potential destinations one can reach in a short time. A fast train that connects to a desirable city sub-center should lead to gentrification along its nodes, and transit improvements (McMillen and McDonald (1998),Glaeser and Kahn (2001) and Baum-Snow et al. (2005)). If such a train is fast enough then it could reduce the demand to live very close to the destination because people can decentralize while still having access to the destination area. Glaeser et al. (2008) documents that poor people live close to slow public transit, while rich people are attracted to fast public transit in centralized cities such as Boston and New York City.

LINE9 significantly reduces travel times within Seoul. To document this fact, we

[2]The Seoul metropolitan government promised this private company 90 percent of the expected profits for the first five years, 80 percent for the next five years, and 70 percent for the last five years.

calculate the travel time between each apartment unit to twenty major destinations before and after it is built, based on the average train's speed of 50km/hour and a walking speed of 4km/hour. Table 1 presents the one way reduction in travel time (measured in hours) to 20 major destinations in Seoul.[3] For example, across our sample, the average apartment resident experienced a reduction in travel time to Gangnam by roughly 5 minutes each way. Those living within a 1 kilometer radius of new transit enjoyed a 14.4 minute reduction in one way travel time to Gangnam. This is a 35% reduction.

While the new train reduces commute times, we do not believe that its effects are large enough to cause important general equilibrium shifts in the entire Seoul housing market. Starting with the work of Sieg et al. (2004) there has been a growing appreciation that local public goods improvements can have general equilibrium effects on a given city. They studied how Clean Air Act regulations sharply reduced pollution in major sections of Los Angeles and this caused a reshuffling of the population such that richer people moved to previously poorer polluted areas of the city. A hedonic researcher who ignores this migration effect would likely over-state the role of clean air improvements as the sole cause of real estate price appreciation. In our setting, we believe that such GE effects are a second order concern. As we discuss in the next section, the new train's treatment area is only a small portion of Seoul.

2.3 The Supply of Housing Close to the New Transit Stations

When the LINE9 was completed, there were 322 residential apartment complexes within a kilometer of the new transit stations. Since the LINE9 opened the owners of these properties began enjoying a capitalization effect that we will estimate below. As we will document in section 5.1, we do not find evidence of a capitalization effect caused by the

[3]Estimated traveling time is based on the assumption that travelers walk to the closest subway station and take a subway. We assume that the walking speed is 4km/hour and the subway speed is 50km/hour.

announcement of the LINE9 construction. Once the new subway opens, nearby land becomes more valuable in these "treated" areas. Thus, real estate developers have incentives to upgrade existing structures and to build new structures. But, Seoul features stringent construction regulation. It takes an average of 33.3 months to build a new apartment complex (Jeon et al, 2010). Redeveloping existing housing entails overcoming many regulatory burdens. For example, each urban housing redevelopment project proposal in Seoul undergoes a nine stage process that includes a strict safety investigation. For the typical redevelopment project completed between the years 2000 and 2015, it took an average of 8.7 years to complete the reconstruction process. Seoul's regulations also require developers to supply a certain proportion of small apartment types. A U.S literature has studied how regulations limits housing supply (see Glaeser et al. (2005)). The same issues arise in South Korea.

While developers face many restrictions in building, they will have a greater incentive to do so if the marginal revenue from building an apartment is higher. The total revenue a developer collects from producing apartments of certain type in a given location is the price per unit multiplied by the units sold. Our ML estimates will provide an estimate of the former. If each developer is a price taker, then facing the non-linear hedonic pricing function (their revenue curve) they will have an incentive to supply new housing that offers greater revenue. Below, we will use data on the new housing supply by developers combined with our CATE estimates to study this.

3 The Empirical Approach

3.1 The OLS Model

Following Kahn (2007), Billings (2011), Zheng and Kahn (2013) and Gibbons and Machin (2005), we begin by estimating average treatment effects of the number 9 on apartment prices using the difference-in-difference framework. For the apartment type i in district j at time t, its price is expressed as follows:

$$Log(Price_{ijt}) = \beta_1 Line9_i + \beta_2 Line9_i \times AFTER_t + \beta_3 X_{ijt} + \mu_j + \lambda_t + \varepsilon_{ijt}, \quad (1)$$

where $Line9_i$ is distance between the apartment type i and the closest LINE9 station and $AFTER_t$ takes one if time period is after the number 9 opened, and 0 otherwise. X_{ijt} is a set of the apartment characteristics except proximity to the number 9, and μ_j and λ_t are the district fixed effect and the quarter fixed effect, respectively. ε_{ijt} is unobservable disturbance.

To specify the treatment area we split the area with G groups, $\mathcal{G} = \{1, ..., G\}$, based on the distance between the LINE9 station and the apartment. For apartment i in district j at time t, its price is expressed as follows:

$$\begin{aligned} Log(Price_{ijt}) = &\sum_{g=1}^{G} \alpha_g \mathbb{I}\{Group_i = g\} + \sum_{g=1}^{G} \beta_g \mathbb{I}\{Group_i = g\} \times AFTER_t \\ &+ \gamma X_{ijt} + \mu_j + \lambda_t + \varepsilon_{ijt}, \end{aligned} \quad (2)$$

where $Group_i$ is the group dummy of apartment i that takes a value in $\mathcal{G} = \{1, ..., G\}$. In the empirical analysis in Section 5, we consider three groups ($G = 3$), where the first group includes the apartments within 1km of transit stations, the second group includes between

1km and 2km from a transit station, and the third group includes all other apartments. For both econometric specifications, we allow serial correlation in ε_{ijt} within the district of "Dong".

For OLS to yield consistent estimates of the average treatment effects, the unobserved error term $(\epsilon_{ij1}, ..., \epsilon_{ijT})$ must be uncorrelated with the "treatment" variables even after we control the observed apartment characteristics and the two sets of fixed effects. If β_2 measured the "average" benefit (in terms of the apartment price) of the travel time reduction by LINE9, this exogeneity assumption would be problematic in the case where the development within the district changes because of LINE9. For example, suppose that as the new stations open, this triggers the opening of new restaurants and stores close to the new stations. In this case, the LINE9 causes both a reduction of commute times, say, to Gangnam and an improvement in local restaurants. Hence, the OLS estimates recovers a total effect, not the partial effect - the local price appreciation associated with the new transit line that is due to the reduction in travel times. We will return to this point and explore suggestive mechanisms under which the new transit stations occur price appreciation in Section 5.5.

We recognize that home prices reflect future expectations of local amenity changes. If home buyers anticipated that the new train would raise future rents, then they may bid more aggressively for houses before the LINE9 opens. In section 5.1 below, we will study trends in real estate prices in the treatment areas and the control areas before the actual opening of the line. We will show that there is little evidence of an anticipation effect in the treated areas [4]. Given that our main interest is in the local amenity value of improved transit, in the Appendix we will present rent regressions that mirror our main

[4]McDonald and Osuji (1995) and Knaap et al. (2001) provide empirical evidences that housing or land price started reacting in advance of when new transit lines opened. However, Gibbons and Machin (2008) argue that impacts of transport improvements are heavily dependent on economic contexts, and that if housing is treated as consumption goods rather than assets, then anticipation effects can be marginal.

parametric specifications. These rent regressions allow us to focus on the annual service flow generated by the LINE9.

3.2 The Machine Learning Approach

Our Machine Learning approach has important economic content. It allows us to disaggregate the average treatment effect associated with new transit access along a high dimensional set of observed attributes. In the presence of significant heterogeneity, the ML approach offers a much more nuanced approach than the conventional hedonic. For those interested in the economic incidence of public policies, this ML approach provides more precise estimates of exactly which incumbent apartment owners are the biggest winners from the city's public goods investment.

Following Rubin (1974), Heckman (1990) and Abadie (2005), we define $Y^0(i,t)$ as the potential outcome that apartment i attains in period t if untreated, and define $Y^1(i,t)$ as the potential outcome that apartment i attains in period t if treated. The treatment effect is $Y^1(i,t) - Y^0(i,t)$. The fundamental problem is that econometrician cannot observe $Y^1(i,t)$ and $Y^0(i,t)$ at the same time. Econometricians observe the realized outcome, $Y(i,t) = Y^0(i,t) \cdot (1 - D(i,t)) + Y^1(i,t) \cdot D(i,t)$, where $D(i,t)$ takes one if treated, and zero otherwise. Due to the missing data problem, it is impossible to identify individual treatment effects, which leads researcher to focus on average treatment effects on the treated under the assumption that the average outcomes conditional on X for the treated and the untreated would have followed similar trends if not exposed to any treatment.[5] As in Heckman et al. (1997), the conditional average treatment effect on the treated is

[5] In the next section, we will present evidence that the treated and the untreated follow a parallel path.

expressed as follows.

$$
\begin{aligned}
E[Y^1(i,1) - Y^0(i,1)|X, D(i,1) = 1] = \\
&\{E[Y(i,1)|X, D(i,1) = 1] - E[Y(i,1)|X, D(i,1) = 0]\} \\
&- \{E[Y(i,0)|X, D(i,1) = 1] - E[Y(i,0)|X, D(i,1) = 0]\}
\end{aligned} \tag{3}
$$

As noted by Abadie (2005), the estimation process is burdensome; four conditional expectations need to be estimated nonparametrically, and the number of observations may not be large enough to estimate conditional expectation when X is high dimensional. To address the innate limitations of DID estimator Abadie (2005) suggests the semiparametric approach, and Athey and Imbens (2006) proposed the generalized identification method that provides entire counterfactual distribution of outcomes that would have been realized both for the treated and the untreated, respectively. In an empirical application, Bajari and Kahn (2005) estimate a hedonic model non-parametrically. However, recent development of supervised machine learning enables researchers to estimate conditional expectations using the regression tree. Athey and Imbens (2015) propose the conditional average treatment effect approach in a context where the unconfoundedness assumption holds. We extend this model to the DID context by incorporating the treatment dummy along with the time dummy as splitting variables in the process of growing a regression tree.

We follow the general supervised machine learning approach to grow our regression tree. (Friedman et al. (2001), Breiman et al. (1984) and Athey and Imbens (2015)[6]). Let first $Q^{is}(\hat{\tau}; \alpha, X, Y^{obs})$ and $Q^{os}(\hat{\tau}; \alpha, X, Y^{obs})$ denote in-sample goodness-of-fit measure

[6]Athey and Imbens (2015) develop five supervised machine learning algorithms for the cases where the unconfoundedness assumption is met. Our approach is based on the single tree with the observed outcome among the five.

and out-of-sample goodness-of-fit measure, respectively, as follows.

$$
\begin{aligned}
Q^{is}(\hat{\tau}; \alpha, X, Y^{obs}) &= -\frac{1}{N}\sum_{i=1}^{N}(Y_i^{obs} - \hat{\tau}(X_i))^2 - \alpha \cdot K \\
Q^{os}(\hat{\tau}; \alpha, X, Y^{obs}) &= -\frac{1}{N}\sum_{i=1}^{N}(Y_i^{obs} - \hat{\tau}(X_i))^2,
\end{aligned}
\tag{4}
$$

where K is the number of leafs in the tree, and α is penalty term to avoid an extremely large tree. $\hat{\tau}(X_i)$ is a sample average of Y_i in leaf. The regularization parameter α is chosen by cross validation, minimizing $Q^{os}(\hat{\tau}; \alpha, X, Y^{obs})$. Let T_M denote a tree with M nodes: $R_1, R_2, \ldots, R_M$. We model the response as a constant $\hat{\tau}(\cdot; T_m)$ in each node and consider the splitting variable j among J explanatory variables and threshold j^{thr} for each region. Using j and j^{thr}, we split parent node(m) into two child nodes(2m and 2m+1).

$$
R_{2t}(j, j^{thr}) = \{X | x_j \leq j^{thr}\} \ \ and \ \ R_{2t+1}(j, j^{thr}) = \{X | x_j > j^{thr}\} \tag{5}
$$

For each $j = 1, \ldots, J$, we fix α and find the value $j^{thr,*}$ that solves

$$
\max_{j^{thr}} \quad Q^{is}(\hat{\tau}(\cdot; T_M^{x_j^{thr}}); \alpha, X, Y^{obs}) \tag{6}
$$

where $T_M^{x_j^{thr}}$ is a new candidate tree generated by splitting the parent node into the children nodes with the threshold of j^{thr}. The following stopping rule is applied;

- If $\max_{j=1}^{J} Q^{is}(\hat{\tau}(\cdot; T_M^{x_j^{thr,*}}); \alpha, X, Y^{obs}) \leq Q^{is}(\hat{\tau}(\cdot; T_M); \alpha, X, Y^{obs})$, then stop splitting and R_m becomes a terminal node
- If $\max_{j=1}^{J} Q^{is}(\hat{\tau}(\cdot; T_M^{x_j^{thr,*}}); \alpha, X, Y^{obs}) > Q^{is}(\hat{\tau}(\cdot; T_M); \alpha, X, Y^{obs})$, then we follow the steps described below.

- If $N_{R_{2m}} < 10$ or $N_{R_{2m+1}} < 10$, stop splitting and then parent node R_m becomes a terminal node where $N_{R_{2m}}$ and $N_{R_{2m+1}}$ are the number of observations in the child node are R_{2m} and R_{2m}, respectively. A very large tree may overfit the data, and it is difficult to interpret average treatment effect within leafs that contain only a single unit (Athey and Imbens (2015))
- If $N_{R_{2m}} \geq 10$ and $N_{R_{2m+1}} \geq 10$, split the node, using variable $j^* = \text{argmax}_j$ $Q^{is}(\hat{\tau}(\cdot; T_M^{x_j^{thr,*}}); \alpha, X, Y^{obs})$ with the threshold of $j^{thr,*}$

- We iterate this process until all of the nodes become terminal nodes and then define T^α as the tree based on the final iteration for a given α, .

In order to choose the optimal penalty parameter, α, we utilize 10-fold cross-validation, minimizing $Q^{os}(\hat{\tau}; \alpha, X, Y^{obs})$. Breiman et al. (1984) prove that a finite number of relative α exist, though possible 'α's are a set of continuous values. This implies that there is the unique T_i that minimizes $Q^{os}(\hat{\tau}(\cdot; \alpha); X^{te}, Y^{te,obs})$ within the interval $[\alpha_i, \alpha_{i+1})$. Taking advantage of the algorithm Breiman et al. (1984) proposed, we construct a sequence of the optimal trees $T(\alpha) =< T_0, T_1, \ldots, T_n >$, corresponding to each relative α_i (See Breiman et al. (1984) for more details). We partition the entire sample into ten subsamples. With only k-1 the training subsamples except the kth subsample, we generate a sequence of tree, $T^k(\alpha)$, using the method described above. With the kth test sample, we estimate the prediction error, using $Q^{os}(\hat{\tau}_{(k)}(\cdot; \alpha); X^{te}, Y^{te,obs})$. For each k, we iterate using the same procedure and find the optimal α^* that solves

$$\arg\max_{\alpha} \ \frac{1}{K} \sum_{k=1}^{K} Q^{os}(\hat{\tau}_{(k)}(\cdot; \alpha); X^{te}, Y^{te,obs}), \quad \text{where} \quad K = 10. \tag{7}$$

With the optimal α^*, we define T^{α^*} and $\hat{\tau}^{\alpha^*}(x)$ to be the optimal tree and the final estimator, respectively.

In the process of growing the regression tree we include the treatment dummy (D) along with the time dummy (T) and other covariates (X) as splitting variables as we did in the linear specification. The treatment dummy equals one if the distance between the apartment and the LINE9 station is less than one kilometer, zero otherwise. Likewise, the time dummy equals one if the year is after the opening of the line and equals zero otherwise. The outcome variable of interest is the log of apartment price (Y), and the covariate vector X includes the apartment size(m^2)[7], the number of rooms, the number of baths, years of depreciation[8], distance to other existing subway transit station[9] and each district dummy. We take advantage of the constructed regression tree, and calculate $E(Y|X = x, D = d, T = t)$ as the conditional expectation. Building on Athey and Imbens (2015), we estimate the conditional average treatment effect(CATE) in DID context as

$$\begin{aligned} CATE =& \{E[Y|X = x, D = 1, T = 1] - E[Y|X = x, D = 0, T = 1]\} \\ & - \{E[Y|X = x, D = 1, T = 0] - E[Y|X = x, D = 0, T = 0]\} \end{aligned} \tag{8}$$

According to Athey and Imbens (2015), our approach represents a single tree model because the treatment dummy, the time dummy and all covariates are included in the single tree. We can extend our approach to the two tree model or the four tree model, based on how splitting variables are included when the tree grows. If the post-treatment effects and the pre-treatment effects are estimated separately from two different trees with subsample of $T = 1$ and $T = 0$, respectively, then it is referred to as the two tree model. In implementing our machine learning approach, the main assumption we are making is

[7]We construct a categorical variable, using 25%, 50% and 75% quantile. One represents the smallest group and four is the largest group

[8]We use a categorical variable that takes on the value of one if less than five years have passed and equals two if between five and ten years have passed, and three otherwise.

[9]We use a dummy that equals one if the distance between the apartment and the existing other station is less than 1km, and equals zero otherwise.

that both the treated and controls would follow a similar trajectory in the absence of the intervention. We discuss pre-trends below to address the issue.

We recognize that the LINE9's geographic placement was not randomly determined. Thus, we are conducting a conditional analysis. Given the line that was built how has it affected real estate pricing? This approach is relevant for an ex-post evaluation of who are the winners and losers of this investment. Our approach cannot be used to predict what will be the future impact of a new Seoul subway in another location.

4 Data

In this section, we describe the data and the summary statistics.

4.1 Apartment Data

We use apartment price data provided by the Ministry of Strategy and Finance of South Korea. This covers more than 90% of entire apartments in South Korea since 2000 and contains a rich set of apartment characteristics, including size, the number of rooms and bath and parking spaces. Since our goal is to investigate the effects of LINE9 on apartment prices, we restrict our sample to the locations where LINE9 passes. In figures 1 and 2, the light green area represents our districts of interest. These data restrictions result in our sample that includes 1,102 apartment complexes and 4,161 apartment types. Market prices are surveyed based on an apartment type rather than at the apartment level. This means that the price data represents the average price of all apartment types that share the same characteristics within the same complex. For example, 84 square meter apartment units with two beds and a bath within the same complex are considered to be the same product

and thus have an identical price in our data.[10] Our data has a panel structure such that the price for each apartment type has been surveyed on a weekly basis. We use the quarterly average price for our analysis. As shown in Table 1, the average apartment has three beds and 1.7 baths, and it is as old as 8.2 years.

4.2 Geographic Information Data

Geographical information data are obtained from the Seoul Metropolitan Government. It provides administrative borders, locations of bus stops and hospital, and all subway systems including LINE9. Figure 2 shows the locations of LINE9 subway stations. Using ArcGIS, we measure the distance between center of each apartment complex and the closest LINE9 station. This is a key variable in our analysis. Our control group consists of apartments more than one kilometer away from the new transit.[11]

As shown in table 2, the mean distance between an apartment and the closest LINE9 station is 1.92 kilometers, and each apartment has other subway station, excluding LINE9, within 0.7 kilometers on average. Our sample of apartments consists of those in districts where the number 9 passes through.

The largest fraction of Seoul residents live in apartments (42.36%, 2014), followed by single-family houses (37.5%). Apartments in Seoul are organized into complexes. The "complex" is composed of several apartment buildings. In our sample, each "complex"

[10] We are interested in the value of a certain apartment type, not individual apartments. An apartment complex has a limited number of apartment types. Under a certain apartment type, there are many homogeneous units. For example, an apartment complex has 600 units, but they can be categorized into four different types, meaning that each type has 150 units on average. We cannot observe transaction prices of individual units but we observe an appraisal of the type. The appraisal process uses each property's selling price and then an averaging takes place. Though our price data has some measurement error due to this process, this is classical error.

[11] In order to control for the direct impact of buses and hospitals on apartment price from benefits from the new subway, we also construct one kilometer buffers for each apartment to count the number of bus stops and hospitals within 1km. These are utilized as controls along with the apartment characteristics.

features an average of 5.5 apartment buildings. Each apartment building contains many apartment units where households reside in. All "apartment units" can be classified into a few number of "apartment type" that share the same apartment characteristics e.g. size, the number of rooms and baths to name but a few. Within the same "apartment type", it is reasonable to assume that apartment units are homogenous. Table 2 shows that each complex has 435 "apartment units" in our sample, meaning that each "apartment building" contains more than 80 "apartment units'".

The Seoul housing stock is quite young. The average age of a Seoul apartment in our data is eight years. It is worth noting that the supply of apartments in Seoul has been expanding since the mid 1970s. The city's growing population and rapid economic development since the 1970s catalyzed the need for high-density residential structures, The development plan for Gangnam caused a massive apartment supply increase in the 1980s. Our data shows the oldest apartment building is 46-year-old, but a large fraction of apartment were built between 1997 and 2007. The recent economic development in the southern part of the city explains why the average age of the housing stock is eight years.

5 Results

5.1 The Pre-Treatment Trend

In conducting a difference in difference study it is important to demonstrate that the pre-trends for the treatment and control groups are not statistically different. In Figure 4, we define the treatment group as the set of apartments within 1 kilometer of LINE9 transit station and the control group is the set of those apartments located more than 2 kilometers away. In figure 5, the treatment group we further refine this set to represent the apartments located within 1/2 of a kilometer of the closest LIEN9 station while the controls are the

same as in figure 4. These figures are based on apartments located in districts where the new number 9 passes through. Both figure 4 and figure 5 show that the pre-trends are parallel, which implies that both the treated and the controls would have followed the similar path in the absence of the intervention. Figure 5 shows that the gap between the treated and controls has been decreasing significantly after LINE9 opened.

Each bar in figure 6 represents the coefficient for interaction between the dummy of within 1 kilometer and the estimated year dummy along with a 90 percent confidence interval. [12] This shows when the LINE9 started affecting apartment prices, and the new transit station effects becomes statistically significant at 10% after 2009 when the LINE9 opened. This implies that the treated and the untreated had experienced a similar path prior to the LINE9, meaning that they would have followed the same trajectory in the absence of the LINE9 even though detailed blueprint was announced at the early period. All pre-trend analyses indicates that the difference in difference approach is suitable to for estimating the impact of the LINE9 on apartment prices.

5.2 OLS Results

Our first set of results builds on earlier work studying the consequences of Seoul's investment in transit infrastructure (see Kim et al. (2005), Cervero and Kang (2011), Bae et al. (2003), Agostini and Palmucci (2008) and Ahlfeldt (2013)).

Table 3 reports results from a standard linear hedonic pricing regression. Controlling for standard structure attributes, the double difference approach indicates that an extra kilometer of distance from LINE9 station is associated with a 1.7% reduction in the home's

[12] We estimate $Log(Price_{ijt}) = \beta_1 Within1km_i + \sum_{Y=2000}^{2015} \beta_Y Within1km_i \times YEAR_Y + \beta_3 X_{ijt} + \mu_j + \lambda_t + \varepsilon_{ijt}$ where $Within1km_i$ takes one if an apartment type i has the new station within 1km, and zero otherwise. $YEAR_Y$ is a year dummy and X_{ijt} is a set of apartment type i's characteristics in district j at time t. μ_j and λ_t are a district fixed effect and a quarter fixed effect, respectively.

price. We further explore these results by including distance to transit dummies. All else equal, properties within 1 kilometer of the transit experience a 4% price appreciation compared to those more than 2km away from the transit (see column 3). Column 4 documents that there is significant heterogeneity in the treatment effects. The district geographic designation called "Dong" is the smallest administrative level, and Seoul has 424 "Dong"s.

5.3 Machine Learning Results

The ML approach yields 142 estimates of the treatment effect. In Figure 9, we present a histogram of these estimates. Based on our estimates, we find that 89 are positive and 53 are negative. One plausible way to interpret the results is that demands for a certain apartment type is low, meaning that there is an apartment type that consumer would like to buy nearby the transit. One explanation is sub-urbanization. Negative leafs are associated with many apartments in Gangnam according to table7 and table 9, which means some residents move out to find a bigger apartment with less congested environment. Table 8 shows that big apartments in suburb area(Gangseo) benefited, while table 9 shows relatively small apartments in Gangnam lost out on LINE9. This suggests that the new subway catalyzes residents to move out to the suburb because traveling costs became smaller due to LINE9. These findings are in accord with urban economics theory that the ability to move at higher speeds encourages suburbanization(Baum-Snow (2007)).

To simplify the presentation of our finding, we sort our estimates of the 142 treatment effects and report the ten largest and smallest CATE estimates. These results are reported in Table 6.

5.4 Developer Responses to the Shifting Real Estate Price Gradient

If we have recovered the true underlying gradient, then the non-linear pricing function sketches out the developer's revenue function for producing different new housing units. Assuming smooth cost functions with respect to apartment size, if developers can earn a large marginal revenue for bundling certain features then they have a profit incentive to build these in. To be specific, developer i supplies apartment type j to maximize the following profit function.

$$\begin{aligned} \max_{j \in J} \Pi_{ij} &= \pi_{ij} + \epsilon_{ij} \\ \pi_{ij} &= R_{ij} - C_{ij}(L, P, B) \\ B &= B(size) \end{aligned} \tag{9}$$

where R_{ij} represents revenue of apartment type j and C_{ij} is a cost function of constructing apartment type j. L indicates required amounts of land, and P is costs related to attaining permits. B denotes building costs that hinge mainly on size, and ϵ_{ij} is a random component. The probability that developer i supplies apartment type j is

$$Pr\{Y_i = j\} = Pr\{max(\Pi_{i1}, \ldots, \Pi_{iJ}) = \Pi_{ij}\} \tag{10}$$

where Y_i indicates an apartment type chosen by developer i. If ϵ_{ij} is independent and identically distributed with Gumbel (type 1 extreme value) distributions[13], then the probability that type j is chosen by developer i is as follows (McFadden et al. (1973)).

$$Pr\{Y_i = j\} = \frac{exp(\pi_{ij})}{\sum_{j=1}^{J} exp(\pi_{ij})} \tag{11}$$

[13]Its cumulative density function is $F(\epsilon_{ij}) = exp(-exp(-\epsilon ij))$

We do not have any cost of construction data, but our ML estimates provide information on the shape of the revenue function. We test whether new construction is positively correlates with our estimates of the revenue function. Potential buyers are more willing to pay for more attractive apartment types. Our CATE estimates provide a proxy for the revenue a developer will receive from selling a given type of apartment. This suggests that the flow of new construction's attributes should be positively correlated with our CATE estimates.

We study this by constructing three histograms. These histograms are based on properties built at three points in time; before 2002, between 2002 and 2012, and after 2012. These three stages can be thought of as the before, middle period and after the construction of LINE9.

The histograms display the share of all housing units as a function of their CATE. We find that units built before 2002 (when developers at that time were unaware of what the future treatment effects of LINE9 would be) build housing units that are symmetrically distributed around zero. In the post-period, the new units built are clustered in the positive CATE estimates. We interpret this as evidence that developers are focusing their efforts on constructing what the market signals is scarce and valued.

5.5 Testing Two Explanations for the Price Appreciation Effects

In this section, we explore two potential reasons for why transit access is associated with rising real estate prices. One explanation is reduced travel time to popular destinations and the other is that new "consumer city" retail and restaurants co-agglomerate near the new train stations. We test each of these by augmenting our linear hedonic regression to include additional explanatory variables and then we test if the capitalization of transit effect changes.

In Table 6, we report results where we return to the parametric hedonic specification reported in equation (1). Across the eight regressions reported in Table 6, we include different combinations of extra control variables to test if the treatment effect shrinks as we control for these variables. We use the distance between each apartment and the LINE9 station in column (1) to (4), and we use the dummy variable that indicates whether an apartment locates within a kilometer from the LINE9 station. We add travel times to 20 key destinations in column (2) and (6), while we control a measure of the new restaurants and the retails co-agglomerated near the new transit lines in column (3) and (7).[14] All controls are included in column (4) and (8). The first four columns do not show that the treatment effect shrinks much as the travel times and the the counts of restaurants and retails are included separately. However, column (4) indicates that the treatment effect shrinks around 20% and becomes no longer statistically different from zero at 10%. This implies that the travel time saving and the "consumer city" rising are leading mechanisms. Column (5) to (8) provides another prospective that reductions in travel time is more influential mechanism behind price appreciation. Apartments within a kilometer from the LINE9 station experience a price premium of 3.36%. Controlling for travel times makes the treatment effect not statistically different from zero at the 10% significance level, while the treatment effect is still statistically significant with controls for retails and restaurants. The findings suggest that the price appreciation is mainly caused by travel time savings rather than by a "consumer city" effect.

We also take our panel CATE estimates from the ML procedure and we compare these leaf specific estimates to those obtained when we conduct a "long difference" ML estimation. In this second case, we only keep the data for the first year and the last year

[14] We use counts of restaurants and retail establishments and the number of employees in those industries at the "Dong" level. "Dong" is the smallest administrative level. The data is drawn from the Seoul metropolitan government

of our sample and we rerun the ML estimator. In Figure 10, we graph the relationship between the long run CATE and the short run CATE. The slope is 0.42. This suggests that the CATE effects shrinks over time. The first possible explanation is that the local "consumer city" effect is small as time passes, which is consistent with what we found in table 6. Another explanation is that there is a general equilibrium effect as developers build new desirable housing units (as revealed by the CATE responses by developers). As the developers engage in this activity, increase in supply lowers the equilibrium prices.

5.6 Estimating the Value of Time

We study what is the implied value of time for different Seoul residents if all of the observed capitalization effect is due to time savings. To study this, we first regress the rent for apartment i in district j at time t on each travel time to twenty major destinations presented in table 1 with apartment type fixed effect as follows.

$$Rent_{ijt} = \sum_{g=1}^{20} \beta_g Hours_{igt} + X_{jt} + \mu_i + \lambda_t + \varepsilon_{ijt}, \tag{12}$$

where $Hours_{igt}$ represents travel time between apartment i and destination g at time t, and X_{jt} includes counts of restaurants and retail shops and the number of employees in those industries in district j at time t. μ_i is apartment type fixed effect and λ_t is quarter fixed effect.

The main reason we use the rent data is to rule out any speculative demand that may affect the property price, and focusing on instant benefits. Though an apartment is not located in the vicinity of the transit station, travel time from the apartment to each destination changed because riders might use a faster route after LINE9 opened. This reveals the correlation between an hour reduction to each destination and rent. We find that tenants are likely to pay US$ 1,454,545 more rent as one travels to Kangnam(CBD) an hour ear-

lier (Table 7, A). Note again that rent is not monthly payments but two-year deposit unlike the U.S. and many countries. This suggests that willingness to pay to be an hour closer to Kangnam is not the amount of deposit per-se, but foregone interest that tenants would have eared if they live in their own apartment. Assuming an interest rate of 2%, the opportunity cost for two years is US$ 29,090(Table 7, B), which means tenants that sacrifice US$ 39 everyday (Table 7, C).

We compare our estimated value of saving an hour in commute time to Gangnam and the taxi fare in Seoul to cross-validate our estimates. Based on the current taxi fares, riders pay US$ 2,73 for first 2km, even though they travel less than 2km. After 2km, riders pay US$ 0.09 for every 142m. With an average speed of 35.4km in Seoul (The Korea Transport Institute, 2011), the estimated taxi fare to travel for an hour is US$ 24.11, which implies that riders pay US$ 48.22 for a round-trip. If one commuted only by a taxi, she would save US$ 48.22 everyday by moving to the region where she is able to travel to Gangnam an hour faster.

6 Conclusion

Over the years 2000 to 2009, US$818 millions were spent to build a new subway in Seoul, South Korea. Such place based investments offer the opportunity to explore how a city's urban form and real estate pricing are affected by such an investment. This paper has used ML methods to contribute to the urban transit infrastructure effects literature.

Our paper implements a difference in difference empirical design. We find that the introduction of the train is associated with apartment price appreciation for certain leafs but actually lowered apartment price growth in other neighborhoods. We posit that the fast train is most likely to reduce prices for apartments in the destination area of Gangnam because people can now decentralize and still access this location by using the fast train.

The notable feature of our study is our ability to document significant heterogeneity on observable dimensions. The payoff for urban research from ML methods is the ability to search across a large number of dimensions of heterogeneity at low cost. Such conditional average treatment effects disaggregate the overall average treatment effect that has been the typical object of interest in earlier real estate studies. By estimating the CATEs our work has new implications for estimating the economic incidence of public transit improvement projects.

In addition to presenting new ML estimates, we have also studied the causes of the positive treatment effects. We find that both commute time reductions to desirable locations and the opening of new stores and restaurants close to the new stations contribute to the price premium. Finally, we have explored how developers of new housing respond to the shifting pricing gradient. Developers produce new units featuring the highest CATE values.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Agostini, C. A. and Palmucci, G. A. (2008). The anticipated capitalisation effect of a new metro line on housing prices. *Fiscal studies*, 29(2):233–256.

Ahlfeldt, G. M. (2013). If we build it, will they pay? predicting property price effects of transport innovations. *Environment and Planning A*, 45(8):1977–1994.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Athey, S. and Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050:5.

Bae, C.-H. C., Jun, M.-J., and Park, H. (2003). The impact of seoul's subway line 5 on residential property values. *Transport policy*, 10(2):85–94.

Bajari, P. and Kahn, M. E. (2005). Estimating housing demand with an application to explaining racial segregation in cities. *Journal of business & economic statistics*, 23(1):20–33.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805.

Baum-Snow, N. and Kahn, M. E. (2000). The effects of new public projects to expand urban rail transit. *Journal of Public Economics*, 77(2):241–263.

Baum-Snow, N., Kahn, M. E., and Voith, R. (2005). Effects of urban rail transit expansions: Evidence from sixteen cities, 1970-2000 [with comment]. *Brookings-Wharton papers on urban affairs*, pages 147–206.

Beygelzimer, A. and Langford, J. (2009). The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM.

Billings, S. B. (2011). Estimating the value of a new transit option. *Regional Science and Urban Economics*, 41(6):525–536.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cervero, R. and Kang, C. D. (2011). Bus rapid transit impacts on land uses and land values in seoul, korea. *Transport Policy*, 18(1):102–116.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Foster, J., Taylor, J., and Ruberg, S. (2010). Subgroup identification from randomized clinical data. *Statistics in Medicine*, 30:2867–2880.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Gibbons, S. and Machin, S. (2005). Valuing rail access using transport innovations. *Journal of urban Economics*, 57(1):148–169.

Gibbons, S. and Machin, S. (2008). Valuing school quality, better transport, and lower crime: evidence from house prices. *oxford review of Economic Policy*, 24(1):99–119.

Glaeser, E. L., Gyourko, J., and Saks, R. (2005). Why is manhattan so expensive? regulation and the rise in housing prices. *The Journal of Law and Economics*, 48(2):331–369.

Glaeser, E. L. and Kahn, M. E. (2001). Decentralized employment and the transformation of the american city. Technical report, National Bureau of Economic Research.

Glaeser, E. L., Kahn, M. E., and Rappaport, J. (2008). Why do the poor live in cities? the role of public transportation. *Journal of urban Economics*, 63(1):1–24.

Gonzalez-Navarro, M. and Turner, M. A. (2016). Subways and urban growth: evidence from earth.

Heckman, J. (1990). Varieties of selection bias. *The American Economic Review*, 80(2):313–318.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Kahn, M. E. (2007). Gentrification trends in new transit-oriented communities: Evidence from 14 cities that expanded and built rail transit systems. *Real Estate Economics*, 35(2):155–182.

Kain, J. F. (1990). Deception in dallas: Strategic misrepresentation in rail transit promotion and evaluation. *Journal of the American Planning Association*, 56(2):184–196.

Kanemoto, Y. (1988). Hedonic prices and the benefits of public projects. *Econometrica: Journal of the Econometric Society*, pages 981–989.

Kim, J., Zhang, M., et al. (2005). Determining transit's impact on seoul commercial land values: An application of spatial econometrics. *International Real Estate Review*, 8(1):1–26.

Knaap, G. J., Ding, C., and Hopkins, L. D. (2001). Do plans matter? the effects of light rail plans on land values in station areas. *Journal of Planning Education and Research*, 21(1):32–39.

McDonald, J. F. and Osuji, C. I. (1995). The effect of anticipated transportation improvement on residential land values. *Regional science and urban economics*, 25(3):261–278.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

McMillen, D. P. and McDonald, J. F. (1998). Suburban subcenters and employment density in metropolitan chicago. *Journal of Urban Economics*, 43(2):157–180.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sieg, H., Smith, V. K., Banzhaf, H. S., and Walsh, R. (2004). Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review*, 45(4):1047–1077.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zheng, S. and Kahn, M. E. (2013). Does government investment in local public goods spur gentrification? evidence from beijing. *Real Estate Economics*, 41(1):1–28.

Figure 1: Map of Seoul

Figure 2: Line 9

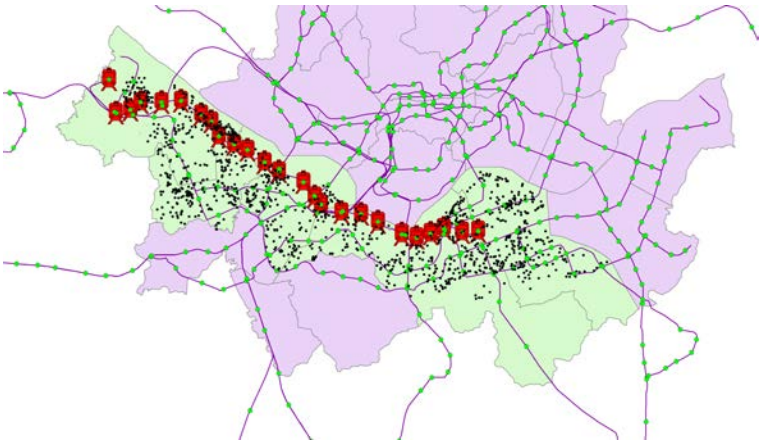


Figure 3: Time line

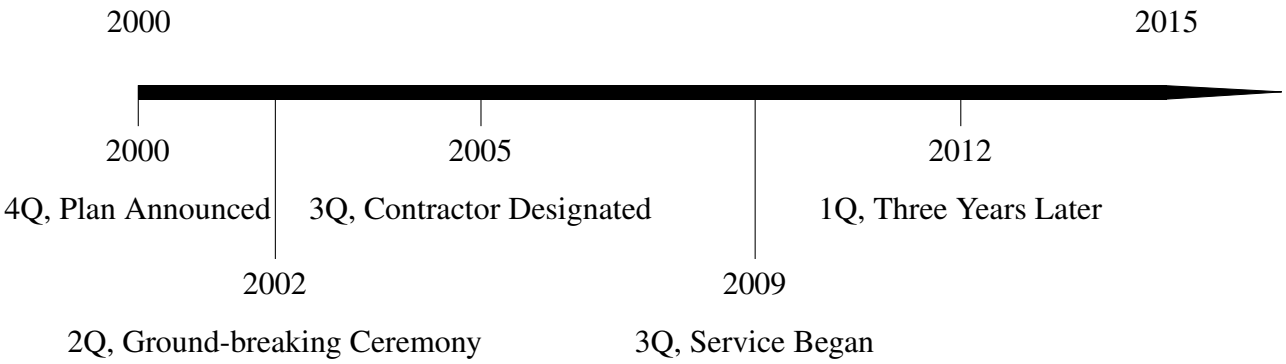


Table 1: Travel Time to Major Destination in Seoul

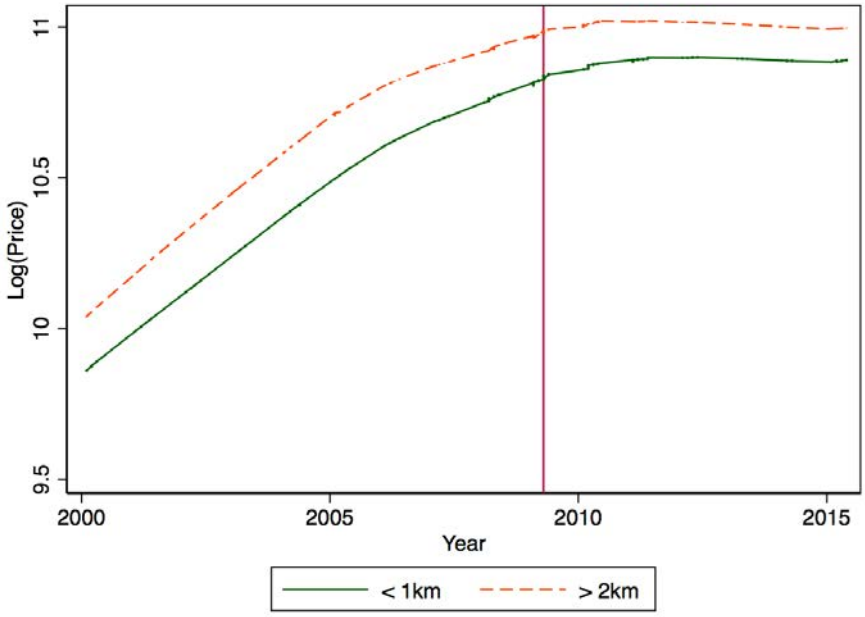
Destinations	Category	Whole Sample			Within 1km		
		Traveling Time before the LINE9 (Hours)	Traveling Time after the LINE9 (Hours)	Difference (Hours)	Traveling Time before the LINE9 (Hours)	Traveling Time after the LINE9 (Hours)	Difference (Hours)
Kangnam	CBD	0.514	0.42	0.094	0.669	0.429	0.24
Yeouido	Business	0.459	0.356	0.103	0.499	0.289	0.21
Myungdong	Business	0.525	0.468	0.057	0.563	0.46	0.103
Hongik University	Entertainment	0.462	0.428	0.034	0.502	0.435	0.067
Express Bus Terminal	Bus Terminal	0.507	0.359	0.148	0.65	0.38	0.27
Shinchon	Entertainment	0.48	0.432	0.048	0.511	0.435	0.076
Sadang	Entertainment	0.674	0.462	0.212	0.71	0.423	0.287
Gimpo Airport	Airport	0.911	0.46	0.451	0.928	0.358	0.57
Incheon Airport	Airport	1.655	1.606	0.049	1.672	1.594	0.078
Seoul Zoo	Entertainment	0.774	0.568	0.206	0.813	0.529	0.284
Gwanghwamun	Business	0.517	0.462	0.055	0.558	0.456	0.102
Kungook University	Entertainment	0.577	0.543	0.034	0.667	0.561	0.106
Nambu Bus Terminal	Bus Terminal	0.511	0.427	0.084	0.67	0.44	0.23
Kangbyun	Bus Terminal	0.593	0.544	0.049	0.701	0.557	0.144
Jongro	Old CBD	0.536	0.483	0.053	0.578	0.478	0.1
Apgujeong	Entertainment	0.534	0.474	0.06	0.654	0.484	0.17
Yeoungdeungpo	Train Station	0.422	0.403	0.019	0.472	0.407	0.065
Seoul Train Station	Train Station	0.49	0.427	0.063	0.528	0.418	0.11
Dongdaemun	Entertainment	0.493	0.466	0.027	0.573	0.485	0.088
Korea University	Entertainment	0.622	0.592	0.03	0.705	0.61	0.095

Notes: Estimated traveling time is based on the assumption that travelers walks to the closest subway station and take a subway, with walking speed of 4km/hour and subway running at 50km/hour.

Table 2: Summary Statistics

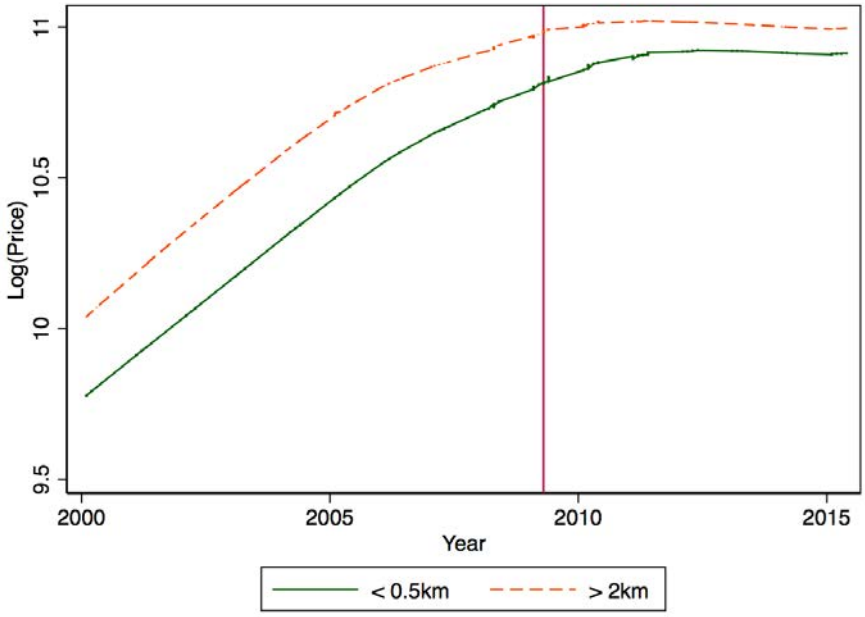
Variable	Mean	Std. Dev.	N
Subway			
Distance to line9 (km)	1.915	1.244	265600
Distance to closest other line (km)	0.702	0.560	265600
Apt Characteristics			
Area (m^2)	94.976	39.501	265600
Room	3.131	0.959	265344
Bath	1.744	0.507	260381
Age, in Years	8.205	9.946	265600
Parking spaces within Complex	546.053	777.993	265600
Number of Apartment units within Complex	435.59	558.62	265600
Number of Apartment buildings within Complex	5.597	9.057	265600
Bus stops within 1km	62.863	19.662	265600
Hospitals within 1km	2.452	1.772	265600

Figure 4: The Pre-Treatment Trend: 1km vs 2km (Lowess Graph)



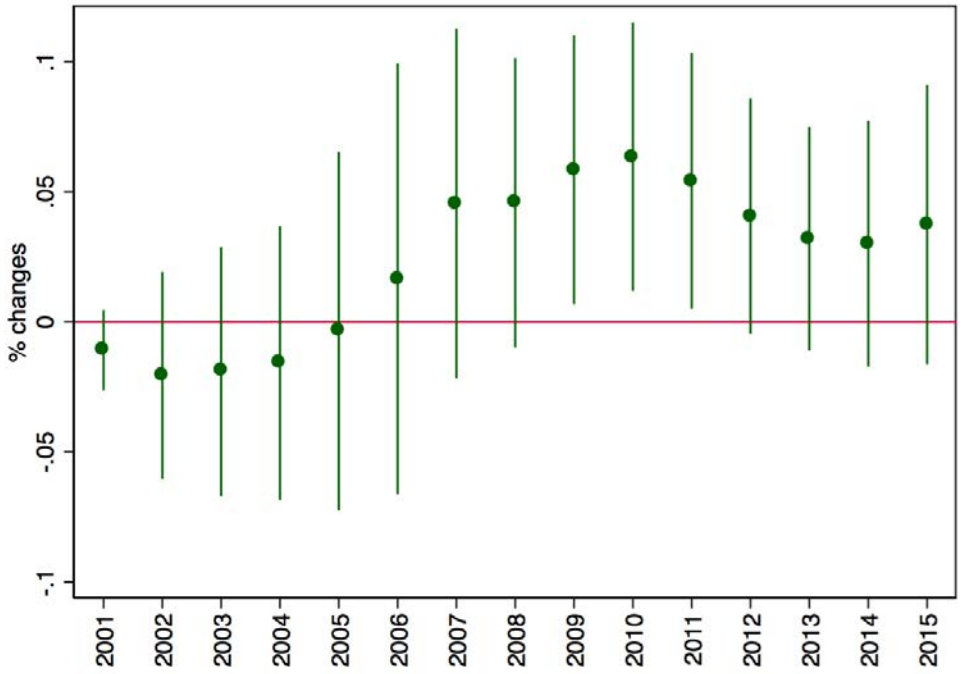
Notes: Vertical line represents when the LINE9 opened.

Figure 5: The Pre-Treatment Trend: 0.5km vs 2km (Lowess Graph)



Notes: Vertical line represents when the LINE9 opened.

Figure 6: Treatment Effect Estimates Over Time



Notes: Each circle indicates the coefficient on the interaction between a "within 1km dummy"and the calendar year. Each bar represents a 90 percent confident interval.

Table 3: OLS Estimates of the Value of Rail Access

VARIABLES	(1) Log(Price)	(2) Log(Price)	(3) Log(Price)	(4) Log(Price)
Distance (km)	0.0339*			
	(0.0193)			
Distance (km) $\times$ AFTER	-0.0174**			
	(0.0078)			
Log(Distance, km)		0.0138		
		(0.0181)		
Log(Distance, km) $\times$ AFTER		-0.0246**		
		(0.0103)		
Within 1km			-0.0785**	-0.0641**
			(0.0367)	(0.0254)
Between 1 $\sim$ 2km			-0.0253	
			(0.0400)	
Within 1km $\times$ AFTER			0.0392*	
			(0.0205)	
Between 1 $\sim$ 2km $\times$ AFTER			0.0224	
			(0.0214)	
Within 1km $\times$ AFTER				0.4073***
				(0.0970)
Within 1km $\times$ AFTER $\times$ Size (m^2) (1)				-0.0017**
				(0.0006)
Within 1km $\times$ AFTER $\times$ Other Line (km) (2)				0.0092
				(0.0101)
Within 1km $\times$ AFTER $\times$ Room (3)				-0.0178
				(0.0285)
Within 1km $\times$ AFTER $\times$ Bath (4)				-0.0269
				(0.0233)
Within 1km $\times$ AFTER $\times$ Age (5)				-0.0117**
				(0.0051)
Within 1km $\times$ AFTER $\times$ Age^2 (6)				0.0001
				(0.0001)
Joint F-value (1) $\sim$ (6)				11.19
(P-value)				(0.000)
Observations	201,985	201,985	201,985	201,985
R-squared	0.9079	0.9076	0.9078	0.9102

Notes: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The standard errors are clustered at the district("Dong") level. Controls include size, the number of room and bath, parking spaces, age, age squared, distance to other closest station, the number of bus stops within 1km, the number of hospitals within 1km, whether it has named brand, the number of households within complex and the number of apartment building within complex. District fixed effect and quarter fixed effect are included

Figure 7: The Regression Tree Result

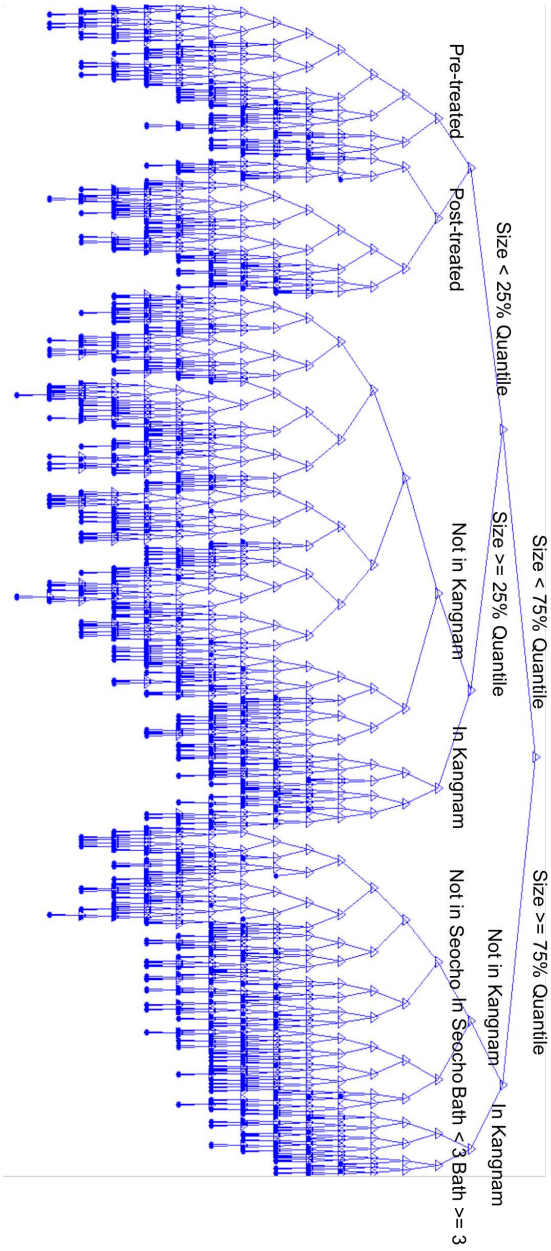


Figure 8: The CATE Distribution

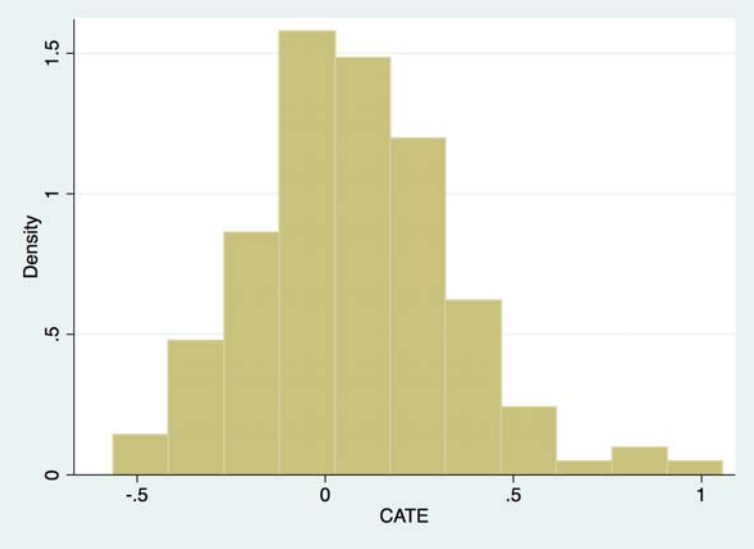


Table 4: Conditional Average Treatment Effects (CATE)

	CATE	N	Size	Room	Bath	Old	Near	County
Bottom 10								
1	-.5619	470	4	4	2	3	1	Kangseo
2	-.5024	120	4	5	2	2	1	Youngdeungpo
3	-.4321	833	1	3	1	3	1	Dongjak
4	-.4003	132	1	2	1	2	0	Yangchun
5	-.3865	784	1	3	1	3	0	Yangchun
6	-.3075	1463	3	3	2	1	1	Kangnam
7	-.3031	628	1	2	1	2	1	Kangnam
8	-.3010	1155	2	3	2	1	1	Kangnam
9	-.2802	282	3	4	2	3	1	Kangseo
10	-.2755	1244	4	4	2	3	1	Dongjak
Top 10								
1	1.0573	201	4	3	2	1	1	Seocho
2	.9023	360	1	1	1	3	1	Youngdeungpo
3	.8203	96	4	5	2	2	0	Yangchun
4	.6403	208	2	3	1	2	1	Seocho
5	.5959	1179	2	3	2	1	1	Seocho
6	.5524	536	3	4	2	1	1	Seocho
7	.5453	787	3	4	2	2	1	Kangnam
8	.5011	1530	2	3	2	2	0	Kangseo
9	.4972	628	4	4	2	2	0	Kangseo
10	.4637	532	1	2	1	2	1	Kangseo

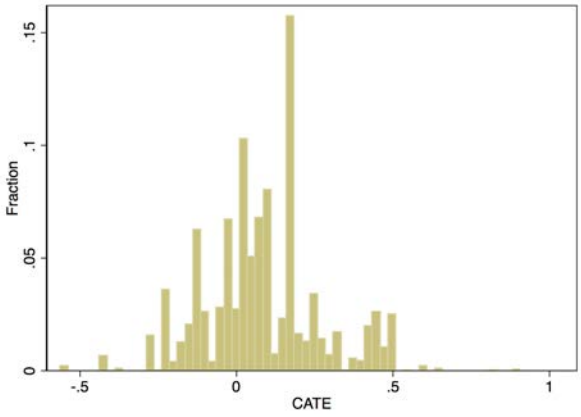
Average	0.08
Number of types with positive impacts	89
Number of types with negative impacts	53
Observations	201,530

Table 5: Composition

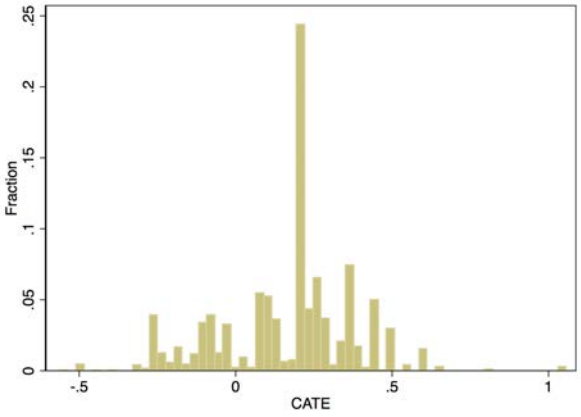
Variables	Top 10%	Top 25%	Bottom 10%	Bottom 25%
Apartment Characteristics				
Size 25	0.178	0.154	0.343	0.347
Size 50	0.323	0.216	0.137	0.214
Size 75	0.286	0.345	0.260	0.236
Size 100	0.213	0.286	0.260	0.203
Room	3.132	3.175	3.288	3.225
Bath	1.80	1.822	1.671	1.700
Less than 5 years	0.273	0.269	0.387	0.274
Between 5 $\sim$ 10 years	0.662	0.611	0.208	0.275
More than 10 years	0.065	0.120	0.405	0.451
Nearby Station (within 1km)	0.555	0.76	0.893	0.982
Districts				
Youngdeunpo	0.065	0.093	0.120	0.162
Dongjak	0	0.063	0.225	0.177
Kangnam	0.086	0.160	0.368	0.276
Kangseo	0.470	0.302	0.135	0.166
Seocho	0.291	0.337	0.046	0.096
Yangchun	0.088	0.045	0.107	0.123

Figure 9: The Empirical Distribution of New Construction as a Function of the CATE

(a) Before 2002



(b) Between 2002 and 2012



(c) After 2012

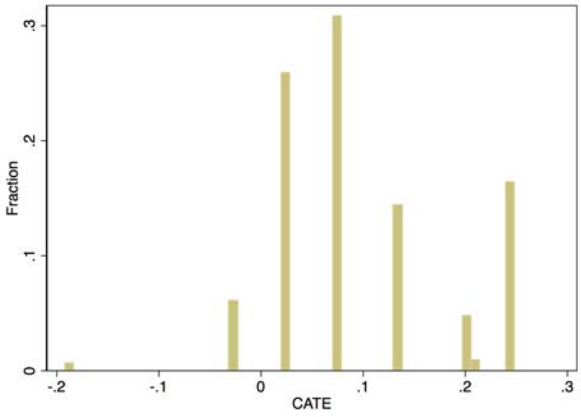


Table 6: Rail Transit Capitalization: The Role of Travel Time Savings and the Local Consumer City

VARIABLES	(1) Log(Price)	(2) Log(Price)	(3) Log(Price)	(4) Log(Price)	(5) Log(Price)	(6) Log(Price)	(7) Log(Price)	(8) Log(Price)
Distance (km)	0.0256	0.0181	0.0251	0.0182				
	(0.0189)	(0.0185)	(0.0187)	(0.0185)				
Distance (km) × AFTER	-0.0172**	-0.0156*	-0.0153*	-0.0138				
	(0.0078)	(0.0090)	(0.0077)	(0.0087)				
Within 1km					-0.0478*	-0.0328	-0.0447*	-0.0314
					(0.0244)	(0.0253)	(0.0244)	(0.0254)
Within 1km × AFTER					0.0336*	0.0248	0.0284*	0.0225
					(0.0173)	(0.0256)	(0.0166)	(0.0256)
Travel Times	No	Yes	No	Yes	No	Yes	No	Yes
Retails and Restaurants	No	No	Yes	Yes	No	No	Yes	Yes
Observations	185,745	185,745	180,569	180,569	185,745	185,745	180,569	180,569

Notes: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The standard errors are clustered at the district("Dong") level. Controls include size, the number of room and bath, parking spaces, age, age squared, distance to other closest station, the number of bus stops within 1km, the number of hospitals within 1km, whether it has named brand, the number of households within complex and the number of apartment building within complex. District fixed effect and quarter fixed effect are included

Figure 10: Long Difference Result

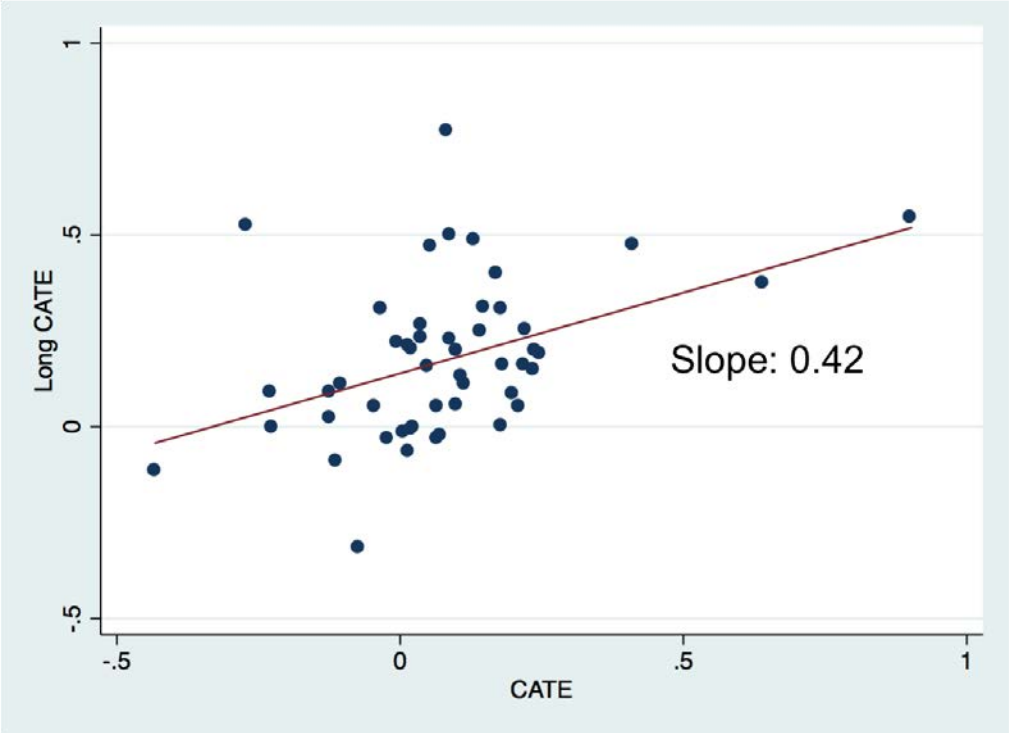


Table 7: Estimates of the Value of Time

The Estimated Value of an Hour	
Correlation b/w An hour reduction in travel time to CBD and 2-year rent deposit	US$ 1,454,545 (A)
Interests for two years (Interest rate 2 %)	US$ 29,090 (B = A $\times$ 0.02)
Daily opportunity costs (1 year = 730 days)	US$ 39 (C = B / 730)
Estimate Taxi Fare for an hour	
Basic Fare (First 2km)	US$ 2.73
Extra Fare	US$0.09 per 142m
Estimated Driving Distance in an Hour (With an average speed of 35.4km/hour, 2011)	35.4km
Estimated Taxi Fare for One-way	US$ 24.11
Estimated Taxi Fare for a Roundtrip	US$ 48.22

Appendix

This section reports an effect of the LINE9 on rent price. Note that the rent here is not monthly payments but two-year deposit unlike the U.S. and many countries. If an anticipation effect or a speculative demand had played a major role in price appreciation due to the LINE9, prices would have experienced a bigger premium than rents. This is because rents are less subject to an anticipation effect. Empirical strategy for the rents are the same as equation (1) and (2), but we replace $Log(Price_{ijt})$ with $Log(Rent_{ijt})$. In comparison to table 3, table A1 reports bigger treatment effects. For every a kilometer closer to the LINE9 station, an apartment experiences of 2.03% premium . If an apartment locates within 1km from the LINE9 station, rents are 6.26% higher than those more than two kilometers away from the new station. Both of them imply that the treatment effect comes from direct benefits like travel time savings or growing retail activities rather than an anticipation of future price hike. Assuming heterogeneous buyers, some who are patient enough to wait for a long time may buy properties in advance and expect a price appreciation later. However, our results show that such cases are not big enough. This justifies that we define the date of opening as a treatment.

Table A1: Impacts of the LINE9 on Rent

VARIABLES	(1) Log(Rent)	(2) Log(Rent)	(3) Log(Rent)	(4) Log(Rent)
Distance (km)	0.0117			
	(0.0181)			
Distance (km) × AFTER	-0.0203***			
	(0.0060)			
Log(Distance, km)		0.0120		
		(0.0138)		
Log(Distance, km) × AFTER		-0.0365***		
		(0.0084)		
Within 1km			-0.0619*	-0.0516*
			(0.0326)	(0.0278)
Between 1 ∼ 2km			-0.0042	
			(0.0282)	
Within 1km × AFTER			0.0626***	
			(0.0194)	
Between 1 ∼ 2km × AFTER			0.0108	
			(0.0197)	
Within 1km × AFTER				0.2555***
				(0.0896)
Within 1km × AFTER × Size (m^2)				-0.0020***
				(0.0005)
Within 1km × AFTER × Other Line (km)				0.0426***
				(0.0127)
Within 1km × AFTER × Room				0.0081
				(0.0208)
Within 1km × AFTER × Bath				0.0024
				(0.0202)
Within 1km × AFTER × Age				-0.0102*
				(0.0053)
Within 1km × AFTER × Age^2				0.0002*
				(0.0001)
Observations	199,742	199,742	199,742	199,742
R-squared	0.9102	0.9105	0.9105	0.9124

Notes: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The standard errors are clustered at the district level. Controls include size, the number of room and bath, parking spaces, age, age squared, distance to other closest station, the number of bus stops within 1km, the number of hospitals within 1km, whether it has named brand, the number of households within complex and the number of apartment building within complex. District fixed effect and quarter fixed effect are included