Data Article

Hedonic dataset of the metropolitan housing market – Cases in South Korea

Yena Song\textsuperscript{a}, Kwangwon Ahn\textsuperscript{b}, Sihyun An\textsuperscript{b}, Hanwool Jang\textsuperscript{b,}\textsuperscript{*}

\textsuperscript{a} Department of Geography, Chonnam National University, Gwangju 61186, South Korea
\textsuperscript{b} Department of Industrial Engineering, Yonsei University, Seoul 03722, South Korea

\textbf{A R T I C L E   I N F O}

\textbf{Article history:}
Received 20 November 2020
Revised 26 January 2021
Accepted 10 February 2021
Available online 24 February 2021

\textbf{JEL classifications:}
R310
R410
R140

\textbf{Keywords:}
Condominium
Economic impact
Metropolitan area
Subway
Transaction price

\textbf{A B S T R A C T}

This article presents a database cleaned and generated for analyzing the economic impact of subway network on housing prices in metropolitan areas. The provision of transit networks and accompanying improvement in accessibility induce various impact and we focused on the economic impact reflected in housing prices. Although our emphasis is on transit accessibility and housing prices, the dataset presented is applicable to other analyses. It includes a wide range of variables closely related to housing prices such as housing properties, local demographic characteristics, local amenities, and seasonal control variables. Various distance variables constructed in a geographic information system environment using public data are useful for exploring the environmental impact on housing prices. These data cover four metropolitan areas—Busan, Daegu, Daejeon, and Gwangju—and provide accurate information on their metropolitan structures distinct from the capital city, Greater Seoul. An empirical analysis performed by Ahn et al. [1] is based on this dataset.

© 2021 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

DOI of original article: 10.1016/j.retrec.2020.100900
* Corresponding author.
\textit{E-mail address:} hwjangg@yonsei.ac.kr (H. Jang).
\textit{Social media:} \textsuperscript{S} (S. An), \textsuperscript{H} (H. Jang)

https://doi.org/10.1016/j.dib.2021.106877
2352-3409/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Specifications Table

| Subject areas | Economics, Geography |
|---------------|----------------------|
| Specific subject areas | Urban Economics/Geography, Econometrics |
| Type of data | Tables |
| How the data were acquired | Official housing land transaction data; distance data calculated in the geographic information system environment |
| Data format | Mixed (raw and pre-processed) |
| Parameters for data collection | Hedonic regression and spatial lag models |
| Description of data collection | Manual and direct data extraction from Statistics Korea, National Spatial Data Infrastructure Portal, and the Korea Transport Database |
| Data source location | South Korea Real Estate Transaction Data: rt.molit.go.kr |
| Kookmin Bank: onland.kbstar.com | Naver Real Estate: land.naver.com |
| Statistics Korea: www.kostat.go.kr | National Spatial Data Infrastructure Portal: www.nsd.go.kr |
| Transport Database: www.ktdb.go.kr | 
| Data accessibility | Repository name: Mendeley Data |
| Direct URL to the data: | https://data.mendeley.com/datasets/d7grg846wv/3 |
| Related research article | Title: Economic impact of being close to subway networks: A case study of Korean metropolitan areas |
| Journal: Research in Transportation Economics | DOI: https://doi.org/10.1016/j.retrec.2020.100900 |

Value of the Data

• A cleaned dataset consisting of condominium transactions in four major of South Korean metropolitan areas was constructed to analyze the impact of subway network proximity on housing prices.
• The dataset can be used to directly calibrate other metropolitan simulation models in which households trade off transportation, central accessibility, and local amenities.
• The data are easily interpreted and can be further processed by quantitative analysis and statistical modeling (e.g., rough-set analysis and hedonic regression).
• Spatial analyses of the geo-referenced data can identify spatial patterns in real estate values.

1. Data Description

Housing prices are closely associated with various factors, and the literature provides a wide range of variables accounting for them. Based on these variables, Ahn et al. [1] analyzed condominium prices in four South Korean metropolitan areas, concentrating on the economic impact of public transit accessibility (specifically, subway accessibility). Aside from the variables that were selected, gathered, and created for that particular purpose, we also included variables that are traditionally considered important to check their statistical associations with condominium prices and avoid omitted-variable bias.

A total of 30 variables were identified, and 26 explanatory variables were divided into four distinct categories: housing properties, local demographic characteristics, local amenities, and seasonal control variables. Table 1 lists the variables in each category along with their measurement scales and coding systems. The cleaned data files for the four metropolitan areas can be accessed at Mendeley Data.1 The spatial data that were used to measure the distances between condominiums and subway networks are provided. Spatial datasets were provided as shapefiles and can therefore be easily integrated into a Geographic Information System (GIS).

---

1 The direct URL to the Mendeley data repository is https://data.mendeley.com/datasets/d7grg846wv/3.
| Variables                        | Measurement scale | Coding system                                                                 |
|---------------------------------|-------------------|-------------------------------------------------------------------------------|
| Condominium price               | Ratio             | Korean won per square meter (won/m²)                                           |
| Date of transaction             | Date              | Date of a condominium transaction, month and year (yyyy-mm)                   |
| Longitude                       | Ratio             | Longitude in the Cartesian coordinate system                                 |
| Latitude                        | Ratio             | Latitude in the Cartesian coordinate system                                   |
| **Housing properties**          |                   |                                                                                |
| Size of unit                    | Ratio             | Expressed in square meters (m²)                                                |
| Floor                           | Interval          | A condominium's floor in the building                                          |
| Number of households            | Ratio             | Number of households in a condominium complex                                 |
| Number of condominium buildings | Ratio             | Number of condominium buildings in a complex                                   |
| Parking spaces per household    | Ratio             | Number of parking spaces ÷ Number of households                                |
| Highest floor                   | Ratio             | Highest floor in a condominium complex                                         |
| Construction year               | Date              | Construction year of a condominium                                             |
| Heating type                    | Nominal (dummy)   | City gas = 0, Others = 1                                                       |
| **Local demographic characteristics** |               |                                                                                |
| Population                      | Ratio             | Number of residents in a neighborhood                                          |
| Population density              | Ratio             | Number of persons per square kilometre (km²)                                  |
| Ratio of adults with higher degrees | Ratio          | (Number of people with a higher degree × 100) ÷ number of people aged 15 or older |
| Young population ratio          | Ratio             | (Number of people younger than 15 × 100) ÷ total population                   |
| Old population ratio            | Ratio             | (Number of people aged 65 or older × 100) ÷ total population                  |
| Medium age                      | Ratio             |                                                                                |
| Sex ratio                       | Ratio             | (Number of males × 100) ÷ Number of females                                   |
| **Local amenities**             |                   |                                                                                |
| Network distance to nearest subway station | Ratio        | Network distance to the nearest subway station, in meters                     |
| Crow-fly distance to nearest subway station | Ratio   | Euclidean distance to the nearest subway station, in meters                   |
| Number of bus stops             | Ratio             | Number of bus stops within a 400-meter radius of a condominium                |
| Network distance to nearest waterfront | Ratio    | Network distance to the nearest river, stream, pond, or seashore, in meters |
| Network distance to nearest greenspace | Ratio   | Network distance to the nearest park, hill, or mountain, in meters            |
| Network distance to the CBD     | Ratio             | Network distance to the city hall, in meters                                  |
| Number of top university entrants | Ratio        | Number of Seoul National University entrants from high schools within a 5-km radius of a condominium |
| Number of high schools          | Ratio             | Number of high schools within 5-km radius of a condominium                    |
| **Control variables for the sales season** |               |                                                                                |
| Spring                          | Nominal (dummy)   | Spring = 1, Others = 0                                                         |
| Fall                            | Nominal (dummy)   | Fall = 1, Others = 0                                                           |
| Winter                          | Nominal (dummy)   | Winter = 1, Others = 0                                                          |

Note: The measurement scales were defined according to the theoretical framework postulated by Stevens [2], who defined four fundamental scales: nominal (qualitative scale characterized by unordered categories); ordinal (expressing qualitative measures by means of ranking); interval (expressing quantitative measures by means of successive values); and ratio (expressing quantitative measures with respect to a true zero point).
environment. These data can be used to develop new spatial variables in their own analytical contexts.

The data presented are useful for investigating the relationship between subway accessibility and housing prices, which is typically expected to be positive. Using our dataset, this relationship can be explored in metropolises operating subway systems with different maturities and shapes. This provides an opportunity to explore and analyze the economic impact of transit networks that are at an early stage on housing prices after controlling for housing-related and environmental factors. In addition, the geo-referenced data enabled the accurate construction of spatial variables. We were therefore able to investigate potential spatial patterns in both housing prices and transit accessibility.

2. Experimental Design and Methodology

The data encompassed four metropolitan areas—Busan, Daegu, Gwangju, and Daejeon—that operate subway systems (see Fig. 1). The presence of a subway system is essential to determining the economic impact of subway accessibility. Although five distinct subway systems operate in South Korea, the system in Greater Seoul was excluded for two reasons. First, the positive economic impact of subway accessibility in this region has been well-documented in previous studies [3–6]. Second, the subway networks in the other four metropolitan areas are less mature, featuring shorter running distances and histories than the very dense, complex, and comprehensive subway network in Greater Seoul [1,7].

We selected condominium prices as a proxy for housing prices. A condominium consists of multiple attached units that are each owned by a separate household. Multi-block complexes with several condominium buildings are also common. Unlike many other countries, South Korea is dominated by condominium housing; in 2018, the condominium household occupancy rate exceeded 50%\(^2\) [8]. The Ministry of Land, Infrastructure, and Transport (MLIT), which gathers and publishes housing transaction prices, provides the most reliable and comprehensive public database. Although the MLIT publishes transaction data for various housing types, these datasets exclude complete location information on each house. Therefore, we cannot create and integrate geographic variables into the dataset. The spatial aspects of our data are crucial because the dependent variable and key factors (more specifically, transit accessibility) inevitably exhibit spatial properties. For example, spatial autocorrelation is commonly noted in housing values, and transit accessibility can be measured by calculating the distance between houses and subway stations. Condominium transaction prices were chosen as the dependent variable for the 2015 period. The price data included all of the transactions conducted in the year 2015.

Data on condominium prices, housing properties, and sales months were obtained from the government with supplements from private real estate companies. Housing properties such as parking space and heating type that are not included in public data were extracted from private databases maintained by Kookmin Bank and Naver Real Estate. We removed observations containing missing data on housing properties, and the number of missing rates for each site was 2.6%, 0.9%, 0.2%, and 0.5% for Busan, Daegu, Gwangju, and Daejeon, respectively.\(^3\) The variables for local demographic characteristics were obtained from Statistics Korea, a public data source.

---

\(^2\) As of 2018, over 50% of households occupied a condominium while approximately 32% occupied single-family homes. With an average condominium occupancy rate of 57%, the figures for each site in the dataset exceeded the national average. The condominium preference is the highest in Gwangju, where more than 65% of households live in condominiums; in the other three metropolitan areas, the condominium occupancy rate is approximately 55%–56%. Single-family homes, the housing type possibly most preferred in many developed countries, account for approximately 30% of the households in each site.

\(^3\) The variables on housing properties—“number of households,” “number of condominium buildings,” “highest floor,” “parking spaces per household,” and “heating type”—include missing data.
Fig. 1. Sites included in the dataset and subway networks in each metropolitan area (a) Site locations and subway maps in (b) Busan, (c) Daegu, (d) Gwangju, and (e) Daejeon. Note: Fig. 1(a) is a modified version of an illustration that appears in [1].
The variables for local amenities are composed of spatial data. In order to create them, we located the condominiums on the map using the longitudes and latitudes in the dataset. This became the key layer for measuring, creating, and testing various environmental variables. Then, transit and road networks and other amenity layers were overlayed by site. By calculating the distance between condominiums and relevant amenities or counting the number of service points within a buffer zone, we constructed various distance variables such as bus accessibility and school-related variables in the GIS environment.

Examples of location-based estimation are illustrated in Figs. 2 and 3. Both figures are based on actual data, but only three condominiums on each map are presented for visual clarity. Fig. 2 depicts the buffering strategy employed in calculating bus accessibility. As presented, each condominium (represented by a dark blue dot) is enclosed by a circular 400-meter buffer radius. The number of bus stops in each buffer zone was counted using the ArcGIS program. In the same manner, the "number of top university entrants" and "number of high schools" were counted in each buffer zone.

Many distance variables were measured as network distance by applying the shortest-path algorithm. However, for subway accessibility measurements, we also included the Euclidean distance, which measures the length of a straight line between a condominium and the nearest subway station. This distance does not reflect the actual travel distance, but works more like the perceptual distance. Fig. 3 presents the calculated distances between the three condominiums and the subway network in the Gwangju metropolitan area, which includes 20 stations. Panels (a) and (b) present the network and Euclidean distances, respectively, calculated using the GIS program. The closest subway station and condominium pairs clearly differ (see Boxes 1 and 3), and the calculated distance between each pair appears different (see Boxes 2 and 4). The "network distance to nearest waterfront," "greenspace," and "CBD" variables were also measured following Panel (a).
Fig. 3. (a) Network and (b) Euclidian distances between condominiums and subway stations in the Gwangju metropolitan area.

Declaration of Competing Interest

The authors have no conflicts of interest to declare.

Acknowledgments

The authors thank our research assistants, Heesoo Kim and Jeehee Oh, for collecting and cleaning the spatial data.

Funding Sources

This research was supported by Yonsei University through the Future-leading Research Initiative (Grant Number: 2020-22-0510; Kwangwon Ahn) and by National Research Foundation
of Korea through the Basic Science Research Program (Grant Number: 2020R1A6A3A01100683; Hanwool Jang).

**Supplementary Materials**

Supplementary material associated with this article can be found in the online version located at doi:10.1016/j.retrec.2020.100900.

**References**

[1] K. Ahn, H. Jang, Y. Song, Economic impacts of being close to subway networks: A case study of Korean metropolitan areas, Res. Transp. Econ. 83 (2020) 100900.
[2] S. Stevens, On the theory of scales of measurement, Science 103 (2684) (1946) 677–680.
[3] C. Bae, M. Jun, H. Park, The impact of Seoul’s subway line 5 on residential property values, Transp. Policy 10 (2003) 85–94.
[4] C. Choi, H. Yun, Spatial impact analysis of the Seoul subway line 7 on apartment properties, Seoul Stud. 5 (4) (2004) 1–12.
[5] J. Kim, M. Zhang, Determining transit’s impact on Seoul commercial land values: An application of spatial econometrics, Int. Real Estate Rev. 8 (1) (2005) 1–26.
[6] K. Lee, K. Kim, Y. Song, Impacts of mixed-use development and transportation on housing values, J. Econ. Geogr. Soc. Korea 13 (4) (2010) 515–528.
[7] H. Kim, Y. Song, An integrated measurement of accessibility and reliability of mass transit systems, Transportation 45 (4) (2018) 1075–1100.
[8] H. Jang, K. Ahn, D. Kim, Y. Song, Detection and prediction of house price bubbles: Evidence from a new city, Lect. Notes Comput. Sci. 10862 (2018) 782–795.