Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 9586307, 15 pages
https://doi.org/10.1155/2023/9586307

Research Article

Time Series Forecasting for Regional Development Composite Index Using Real-Time Floating Population Data

Jungyeol Hong,[1] **Jieun Na**,[1] **Youjeong Kang**,[1] **and Dongho Kim**[2]

[1]*Department of Transportation Engineering, Keimyung University, Daegu, Republic of Korea*
[2]*Department of Big Data Platform and Data Economy Research, Korea Transport Institute, Sejong-si, Republic of Korea*

Correspondence should be addressed to Jungyeol Hong; jyhongsun@gmail.com

Received 3 November 2022; Revised 21 May 2023; Accepted 5 July 2023; Published 4 August 2023

Academic Editor: Rui Jiang

Copyright © 2023 Jungyeol Hong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Composite development indices show an exponential movement of major economic indicators to identify and predict the overall trend of the national economy. However, the existing method of writing composite development indices is based on simple statistical methods using macroscopic data. Therefore, it presents limitations when grasping regional economic trends late. It is because the time of announcement of composite development indices is concentrated at the end of each month, quarter, and year. This study used the floating population estimated from smartphone data that can be collected in real-time to analyze how floating population patterns affect regional economic situations to compensate for these limitations. The primary purpose was to present a prompt development prediction methodology that reflects this meaningful relationship. A correlation and cross-correlation analysis was performed to exhibit a clear relationship between composite development indices and floating population value. In addition, a time series model and a multiple regression model analyses were applied to predict regional development indices. The results obtained facilitated the prompt selection of regional composite indices after choosing a model that exhibits high prediction accuracy and efficiency of the application. The selected regional development composite indices are expected to be used as a faster and more reliable prediction criterion than the existing development composite indices used to predict a specific city's economic situation.

1. Introduction

Due to the long-term spread of COVID-19, Korea has been facing various problems, such as the deterioration of the working-class economy, business stagnation, and the intensification of unemployment. Therefore, there is an urgent need to create prompt development composite indices that allow immediate diagnoses of regional economies to realize timely identification and solving of these urban problems. Development composite indices published by Statistics Korea are created by synthesizing monthly changes in economy-sensitive indicators by the economic sector (productivity, investment, employment, and consumption).

These composite development indices help track the past and present economic trends and can be used as barometers to predict future economic situations. However, Park et al. [1] mentioned the disadvantages of the development of composite indices based on something other than statistical models. They highlighted their inability to formulate correct economic policies owing to their lack of visibility and the difficulty they present in grasping regional economic trends since their time of the announcement is concentrated at the end of each month, quarter, and year.

Therefore, the purpose of this study is to create prompt development composite indices that can predict and indicate, at a glance, near-future economic situations using big traffic data to compensate for the limitations of the existing method of creation of development composite indices. In order to create prompt development composite indices, floating population data from Ulsan Metropolitan City in South Korea were collected, extracted, and processed from smartphone data during the years 2019 and 2020. The advantage of floating population data based on communication data in terms of analyzing future economic trends is that

it contains various types of data such as floating population statistics, origin and destination names, and travel purposes in microtime units such as hourly and daily. In addition, it also allows the collection of data of users regardless of the type of transportation (car, bus, subway, bike, and foot), thereby significantly increasing the sample size and range of applications. Thus, several studies have been conducted on the creation of regional economic indicators using various types of big data (such as call detailed recorder (CDR)) similar to communication data or the creation of indicators that allow the prediction of regional economic activities using ship traffic big data based on automatic identification system (AIS).

However, studies on developing or predicting development composite indices using various types of big traffic data still need to be available in Korea. Thus, to compensate for domestic research conditions, this study aimed to establish a clear relationship between floating populations (real-time traffic big data) and composite development indices and create prompt development composite indices based on statistical models by preparing an estimation methodology framework for Ulsan Metropolitan City.

2. Literature Review

2.1. Study on the Use of Smartphone Big Data. CDRs are digital data recorders used for cell phones and other communications and contain internet usage, payment history, location data, and other communication transactions (i.e., text messages) sent through devices or during calls. Therefore, it refers to all data recorded during cellphone usage that can be used to accurately identify the time, pattern, and location where the user is using his/her mobile phone. The study by Šćepanović et al. [2] analyzed regional call patterns and mobility per user group to estimate the house (place with more frequent calls) and workplace location of users based on the frequency of calls during the day. In addition, time-specific call patterns were used to draw a correlation with the region's energy system, power plant, and energy grid data and thus reveal the presence or absence of gaps between rich and poor, economic activities, and service infrastructures. They concluded that the use of communication data could exhibit a diversity of socioeconomic insights. Arhipova et al. [3] grouped regions and users with similar economic activities based on the number of incoming and outgoing calls and text messages obtained by collecting CDR data every 15 min. Other studies by Arhipova et al. [4, 5] used mobile phone activity data to predict changes in people's behavior and socioeconomic indicators in municipalities due to COVID-19.

Consequently, through the utility curve of 8 individual groups showing different economic activity patterns, it was concluded that the efficiency of the development strategies selected by each municipality could be evaluated. In addition, they proposed a method to create real-time and periodic regional development indices using the amount of phone call activity. Pappalardo et al. [6], based on users' areas of activity, revealed that mobility diversity (MD) is the characteristic that has the most significant impact on economic development. Blumenstock [7] discovered that digital footprints created by users could exhibit the characteristics of specific social groups. Furthermore, Kreiendler and Miyauchi [8] reported that regional economies could be predicted by analyzing commuting patterns by tracking work and residence locations using CDR data. Dong et al. [9] showed that the level of industrial development of each land by use could be evaluated by comparing the number of times that a specific place has been searched on Baidu's map and the actual pedestrian volume of the same place. The existing literature regarding research conducted using big smartphone data is summarized in Table 1.

2.2. Research on the Use of Traffic Big Data. Traffic big data are generally collected using road traffic sensors, loop detectors, and highway Hi-Pass data collectors. As mentioned in the previous section, studies have attempted to create regional development indicators using CDR data, while certain scholars have attempted to predict local economies using different types of traffic big data. For example, Van Ruth [10] introduced a U.S. transportation service index that combined two separate indices (cargo transport and passenger transport). As this transportation service index appeared to have a clear and strong correlation with the U.S. economic activity, it was claimed that Nederland's national road traffic indices might also be excellent business cycle indicators for the nation's economy. In addition, the correlation between local traffic intensity and local manufacturing indicators created based on local traffic data was found to be favorable (0.56), thereby confirming the feasibility of using regional traffic intensity indicators to predict local economies. As presented in Table 2, Arslanalp et al. [11] attempted to predict economic changes using vessel traffic data collected through automatic vessel identification systems. The real-time transaction flow was predicted by developing a "Cargo Number" indicator to calculate the number of vessels and a "Carrying Freight" indicator showing cargo load changes. Since the high correlation between these two indicators, it was proven that real-time data collected through automatic vessel identification systems could improve the timeliness of official trade statistics and aid in identifying turning points in the business cycle. Furthermore, Rowland [12] developed monthly and quarterly production indicators by collecting data regarding the number of visits and anchorages of vessels using automatic vessel identification systems. The developed indicators can identify the level of international trade between goods and economic recessions.

3. Analytical Methodology

3.1. Analysis Overview. The floating population data have information such as the date, day of the week, origin-destination code, age group, gender, the purpose of walking, and the number of floating populations. The walking purposes mainly used in this study include commerce, housing, work, shopping, tourism, and leisure, and the basic statistical analysis of the number of floating populations is

Table 1: List of literature reviews in web/mobile application for the economy prediction.

Authors	Data	Contents	Methods
Šćepanović et al. [2]	CDR (frequency of calls and duration of calls)	(i) Development of a universal model for travel patterns (ii) Analysis of the availability of new socioeconomic indicators that do not depend on local agencies or government	(i) Identifying the location of home and workplace with CDR (ii) Analysis of individual travel and call time patterns (iii) Correlation analysis with the energy grid
Pappalardo et al. [6]	CDR (timestamp, coordinates, and caller, callee)	(i) Mobility diversity is the variable that has the greatest influence on economic development estimation among individual mobility (mobility volume, mobility diversity, social volume, and social diversity)	(i) Configuration of the user's time-resolved trajectory using CDR data and analysis of individual travel characteristics
Dong et al. [9]	Baidumap (anonymous ID, coordinates, timestamp, and query)	(i) Evaluation of the degree of industrial development in the store, company, and service sectors	(i) Using data on the number of people searching for a specific place and the number of pedestrians (ii) Correlation analysis between floating population and industrial development in a specific place
Blumenstock [7]	CDR (phone calls, text message, airtime purchase, mobile money use)	(i) Digital footprint can be used to infer individual's socioeconomic characteristic	(i) Sample extraction by geographically stratifying among mobile phone subscribers (ii) Correlation analysis between telephone survey results and actual property level
Arhipova et al. [3]	CDR (number of outgoing and incoming calls and sent and received calls)	(i) Predicting changes in economic activity in a specific region in real time (ii) Prompt regional comprehensive economic index	(i) Mobile phone call volume in CDR data is a reliable tool for continuous and dynamic monitoring (ii) Used to develop indicators of regional economic activity
Arhipova et al. [4]	Mobile phone activity data and socioeconomic indicators	(i) Analyzing the behavior of people changed in the COVID-19 pandemic (ii) Predicting behavioral changes affected the economic activity in municipalities, taking into consideration significant changes in people's habits and employment conditions	(i) Using the developed regional planning methodology based on the mobile phone activity data and socioeconomic indicators
Arhipova et al. [5]	Mobile phone activity data (the number of outgoing and incoming calls and text messages)	(i) Analyzing how economic activities can improve regional development planning	(i) Examining the economic activity level in each Latvia's municipality in comparison to the mobile phone activity
Kreindler and Miyauchi [8]	CDR (user ID, timestamp, and cell tower location)	(i) Using the commuting pattern between regions, predicting the local economy	(i) Establishment of a commuting matrix by grasping the location of the house and workplace (ii) Estimate commuting costs and analyze the economic status of Google Maps' travel time

Table 2: List of literature reviews in traffic data.

Authors	Data	Contents	Methods
Van Ruth [10]	Local traffic volume by vehicle type and road type (number of vehicles)	(i) Preparation of traffic intensity indicators based on local traffic volume data (ii) Application as a useful indicator for predicting the regional economic index	(i) Correlation analysis between traffic volume index and production and development index collected from road sensors
Arslanalp et al. [11]	Marine traffic big data collected based on the automatic identification system	(i) Indicators of marine trade and indicators of global trade patterns	(i) Development of two indicators: the number of marine traffic index and cargo load index
Rowland [12]	Marine traffic big data collected based on the automatic identification system	(i) Development of monthly and quarterly indicators (ii) The economic downturn and the confirmation of the level of imports of international trade between goods	(i) Monthly indicators of the time the ship stayed at the port and the frequency of visits obtained from the automatic identification system

shown in Table 3. The raw data of the floating population is divided into daily and district units and used as a total floating population every month. As shown in Table 4, the total number of floating population data is 59,345,986, and the number of floating populations for each district is 34.02 per day and 84,128,557 per month.

First, Ulsan Metropolitan City's floating population data were segmented into six trip purposes (commercial, business, residential, tourism, leisure, and shopping) and two travel distances (short and long) to obtain more microscopic and consistent analysis results. Dividing the floating population was defined as short-distance trips when both the origin and destination of traffic occurred in Ulsan Metropolitan City and long-distance trips when one of the origins and destinations was outside Ulsan Metropolitan City. In addition, the floating population was divided into 12 categories based on the purpose of the trip, such as commerce, work, and shopping information included in the floating population data. When the floating population is divided and analyzed in this way, it is possible to understand how the number of floating populations generated by traffic distance and purpose of traffic has a relationship with the regional economic index and the heterogeneous effects of each.

The composite economic index utilizes indicators such as production, consumption, and employment status that reflect the economic situation well and consists of the leading economic index, the accompanying economic index, and the trailing economic index. The leading economic index changes before the actual economic cycle, so it is an index that can predict future economic trends, and the accompanying economic index diagnoses current economic trends as it fluctuates with the current economic cycle. Finally, since the economic follow-up index fluctuates after the current economic cycle, economic trends are defined as an index that can be judged afterward.

As these data are time-sequential, changing over time, trending, irregular, and seasonal factors were removed using statistical techniques after diagnosing its trend and seasonality. It eliminated the possibility of obtaining results biased by certain factors. Subsequently, the composite economic indices with a high correlation with the aforementioned 12 floating population data categories were created by performing a correlation and cross-correlation analysis. The obtained regional development composite economic indices were defined as dependent variables. Each floating population category and COVID-19 variables were considered independent variables to develop multiple regression analysis models and a SARIMAX (seasonal autoregressive integrated moving average with exogenous variables) time series model. Finally, the possibility of offering prompt development composite indices using floating population data was presented by selecting a model with excellent predictive performance for local economies. Figure 1 is a flow chart that the research process of analyzing and predicting prompt the composite economy using floating population data estimated based on communication big data.

3.2. Correlation and Cross-Correlation Analysis. The correlation analysis shows the correlation between two continuous variables as a correlation coefficient between −1 and 1. Therefore, the analysis used to measure linear relationships was used to diagnose the degree of relation between regional composite indices and floating population data categories. In this study, based on the correlation coefficient range presented in the study by Dancey and Reidy [13], the linear correlation of two variables was defined as perfect (correlation coefficient: 1), strong (correlation coefficient: 0.7–0.9), moderate (correlation coefficient: 0.4–0.6), and weak (correlation coefficient: 0.1–0.3). Therefore, we aim to select floating population data and development composite index combination with a "Strong" correlation coefficient.

In addition, a cross-correlation analysis was performed on indicators that showed a high correlation with floating population data to confirm their correlation. Cross-correlation analysis determines precedence or antecedence between two variables by determining their correlation with time differences based on their size, sign, and significance [14]. In other words, it can judge variables' precedence, antecedence, and accompaniment because they are not independent of time. Before model building, the purpose was to find indicators whose correlation analysis coefficient and the average value between this coefficient and the cross-correlation analysis coefficient were above the "Strong" range (0.70). As cross-correlation analyses are performed to show the linear relationship and similarity of two-time series data, the degree of correlation between the value of an observed variable x at a point in time, t, and the value of an observed variable y at a point in time, $t + k$, can be confirmed. They are performing the cross-correlation analysis, values when $k = 0$ are referred to as cross-correlation coefficients, while those when $k \neq 0$ are referred to as lag correlation coefficients. Based on the results of equations (1) and (2), the moment when the time difference k, which is the value with the highest correlation coefficient, equals 0, is a comovement indicator wherein two variables change simultaneously. In contrast, the moment when the time difference k is negative is a precedence indicator wherein the observed variable x progresses first, followed by a change in the value of y.

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \overline{x})(y_{t+k} - \overline{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \overline{x})^2 (y_t - \overline{y})^2}}, \ (\mathrm{k} = 0, \ \pm 1, \ \pm 2), \tag{1}$$

where x_t is the value of the observed variable x at a moment t, y_{t+k} is the value of the observed variable y at a moment $t + k$, $\overline{x}$ is the average value of the sample group x, and $\overline{y}$ is the average value of the sample group y.

$$\text{When k} = 0, r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \times \sum(y_i - \overline{y})^2}}. \tag{2}$$

Here, x_i is the observed i value of the x sample group, $\overline{x}$ is the average of the x sample group, y_i is the observed i value of the y sample group, and $\overline{y}$ is the average of the y sample group.

Table 3: Structure of floating population data.

STD_YMD	Day	District_Gu_Code	...	Dst_HCODE	Age	Sex	Type	Pop
20190101	Tue	11110	...	3114056000	20G	2	BZ	2.505495
20190101	Tue	11110	...	3114056000	60G	1	BZ	1.904348
20190101	Tue	11110	...	3114062500	60G	1	NO	1.564626
20190101	Tue	11110	...	3114059500	10G	1	WK	4.408163
...	...	...	...	...	...	...	...	...

Table 4: Descriptive statistics of floating population data.

Category		Value	
		Aggregate data by day	Aggregate data by month
Number of data		59,345,986	59,345,986
Floating population (ped/day or Ped/month)	Sum	2,019,085,364	2,019,085,364
	Mean	34.02	84,128,557
	Max	15,238.42	90,506,632
	Min	0.012	75,319,979
	Standard deviation	286.929	4,475,623
Temporal range		Day unit	Month unit
Spatial range		District unit	District unit

3.3. Time Series Model. Time series analysis is a method that estimates future values by separating changes in one variable into a long-term trend, periodic, seasonal, and irregular changes over time. Most development composite indices use time series analysis methods to remove noneconomic factors (seasonal and irregular factors). We collected monthly development composite indices and floating population category data of Ulsan Metropolitan City from 2019 to 2020 as time series data. Therefore, similar to the development of composite indices, data stationarity was secured by differencing the change factors floating population data to secure the reliability and consistency of analyses between two variables. Seasonal autoregressive integrated moving average (SARIMA), which is a time series model, is a method that compensates for the shortcoming of the autoregressive integrated moving average (ARIMA) model. However, it has poor prediction accuracy when analyzing data with seasonal or periodic characteristics, owing to an inability to efficiently reflect the periodic characteristics of time series data. The estimation process of SARIMA proceeds in the following order:

(i) Identification (verification of the presence or absence of seasonality and confirmation of the stationarity of time series)

(ii) Estimation (estimation of the most suitable p, d, q, P, D, and Q values)

(iii) Estimation of the final model selected through diagnosis (residual analysis and overfitting diagnosis)

SARIMA is defined as shown in the following equation:

$$\begin{aligned} \text{SARIMA} &= \phi_P(L)\Phi_P\left(L^s\right)(1-L)^d(1-L)^D Z_t \\ &= \delta + \theta_q(L)\Theta_Q\left(L^S\right)\epsilon_t. \end{aligned} \tag{3}$$

Here, Z_t represents the native series data, D represents the degree of seasonal disparity, t represents the time operator, p represents the AR term disparity, ϵ_t represents the error terms and white noise following $N(0, \sigma^2)$, q represents the MA term disparity, L represents the retrograde operator, d represents the difference disparity, P represents the SAR term disparity, δrepresents the constant, and Q represents the SAM term disparity.

Regional development composite indices can be affected by time serial and floating population factors, which are considered related in this study. In addition, they can also be affected by exogenous factors such as COVID-19, strikes, and economic recessions. As these exogenous variables are considered to wither the growth of developing composite indices, the seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX) model, which considers exogenous variables, was used to reflect exogenous factors in dependent variables [15]. In the SARIMAX model, exogenous variables compensate for the limitations of univariate time series that only used past time series data. This method increases the explainability of exogenous variables and the predictability of dependent variables more than existing models. It can be expressed as shown in equation (5):

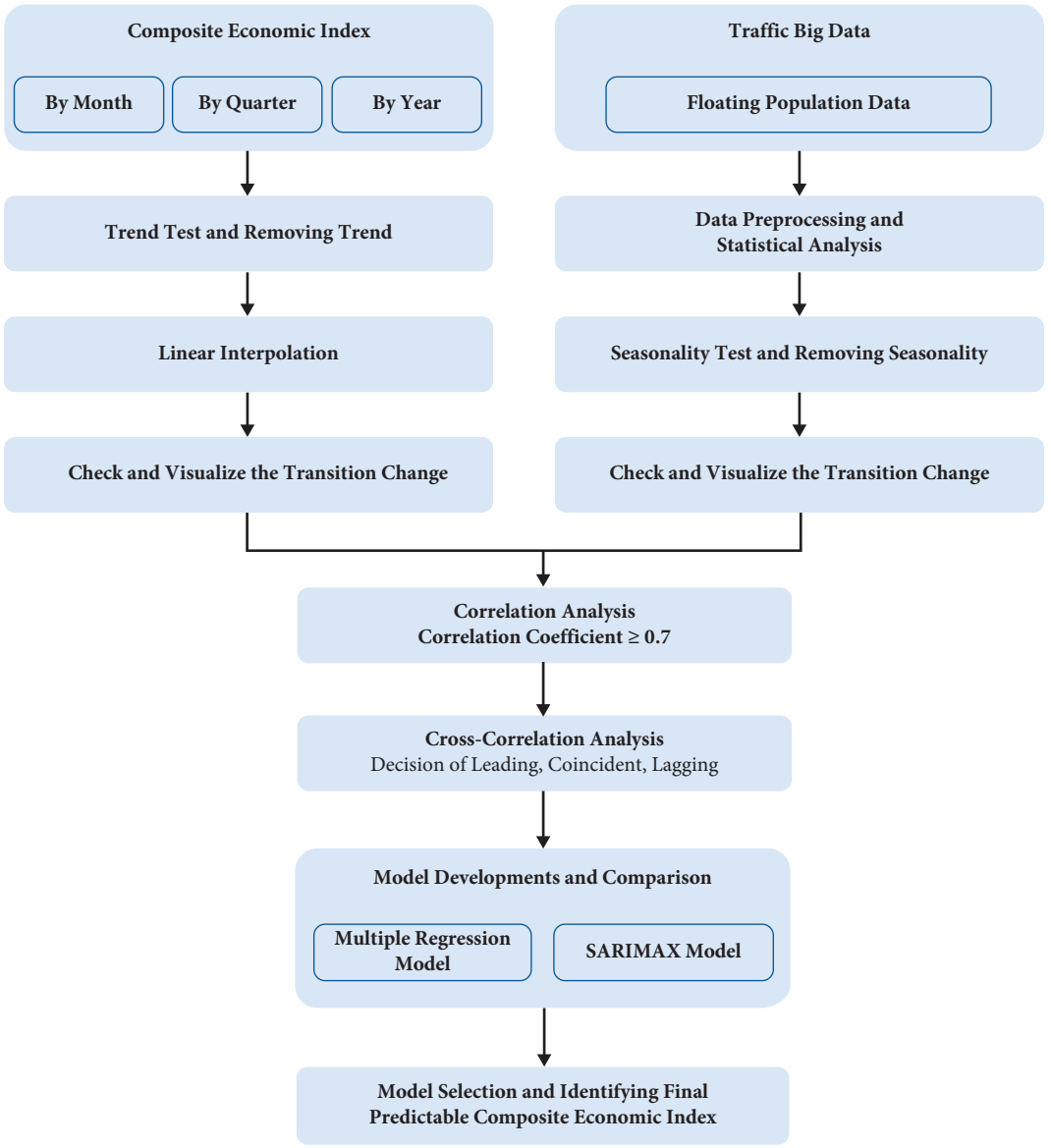

Figure 1: Analysis framework for the study.

$$w_t = y_t - \beta_1 x_{1,t} - \beta_2 x_{2,t} \cdots - \beta_b x_{b,t}, \tag{4}$$

$$\text{SARIMAX} = \left(1 - \sum_{i=1}^{p} \varnothing_i L^i\right)\left(1 - \sum_{j=1}^{p} \Phi_j L^{j\times s}\right)(1-l)^d (1-L^s)^D w_t - \eta = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\left(1 + \sum_{j=1}^{Q} \Theta_j L^{j\times s}\right)\alpha_t, \tag{5}$$

where L represents the time difference operator, y_t represents the observation data at a t point in time, $x_{k,t}$ represents the kth exogenous input variable at a t point in time, β_k represents the coefficient value of the k^{th} exogenous input variable, b represents the number of exogenous variables, w_t represents the autoregressive error, p represents the nonseasonal AR disparity, P represents the seasonal AR disparity, q represents the nonseasonal MA disparity, Q represents the seasonal MA disparity, d represents the number of nonseasonal time series differences, D represents the number of seasonal time series differences, η represents the constant of the SARIMA model, and α_t represents the error of the time difference t.

4. Analysis Results

4.1. Correlation and Cross-Correlation Analysis Results. Table 5 presents the results of the correlation analysis between the floating population categories and development composite indices. Service industry (transportation, warehousing, lodging, restaurants, arts, sports, and leisure) production indices also exhibited a high correlation (0.75)

Table 5: Correlation coefficient between floating population with specific trip purpose and composite economic indexes.

Composite economic index	Wholesale and retail	Transportation and warehousing industries	Accommodation and restaurant business	Education service industry	Arts, sports, and leisure services	Associations and organizations repair and other personal service industry	Professional retail store
Leisure	0.49	0.70	0.86	0.38	0.80	0.52	0.57
Leisure with short distance	0.72	0.78	0.81	0.35	0.80	0.74	0.71
Commercial	0.55	0.80	0.93	0.58	0.91	0.76	0.81
Commercial with short distance	0.64	0.77	0.88	0.46	0.85	0.82	0.83
Commercial with long distance	0.02	0.61	0.80	0.86	0.86	0.27	0.42
Tour	0.67	0.71	0.82	0.34	0.80	0.70	0.59
Tour with long distance	0.58	0.70	0.82	0.38	0.80	0.64	0.49
Work	0.76	0.77	0.80	0.22	0.75	0.87	0.72
Work with short distance	0.81	0.67	0.68	0.07	0.61	0.88	0.72
Work with long distance	0.16	0.74	0.83	0.65	0.85	0.36	0.33
Shopping	0.54	0.85	0.94	0.43	0.88	0.71	0.68
Shopping with short distance	0.55	0.85	0.94	0.44	0.88	0.71	0.69
Shopping with long distance	0.44	0.80	0.90	0.34	0.80	0.64	0.56

with the floating population. In particular, the lodging and restaurant business production indices exhibited a very high correlation (more than 0.93) with floating populations related to commercial and floating purposes. Lodging refers to accommodations, campsites, and camping facilities, and restaurant businesses refer to dining restaurants, cafes, pubs, and independent dining cars. Because most of them are small businesses, they are sensitive to local economic booms and recessions, and the amount of floating population with commercial and shopping purposes is proportionally related to the activation of service industries such as lodging and restaurant businesses.

Moreover, transport and warehousing industries are related to the movement of people, transportation (by land, water, and air), and storage of goods. The floating population tends to increase usually during the holiday or economic booms, thus closely affecting the local economic situation. Therefore, it can be assumed that the leisure and shopping floating population value is related to the activation of transportation and warehousing industries. Art, sports, and leisure services refer to multiple leisure industries such as movie theaters, musicals, plays, museums, zoos, golf courses, and water activities. The more the average income of people increases and the local economy improves, the more the leisure transit increases.

Cross-correlation analysis can be interpreted as causal relationships by analyzing precedence and antecedence factors according to the time difference. Therefore, a cross-correlation analysis was performed to determine the time difference relationship between the floating population and the incremental patterns of development composite indices. Cross-correlation analyses analyze the precedence, comovement, and antecedence using a k value that refers to time differences. Precedence ($k < 0$) refers to when floating population changes occur before economic changes. Furthermore, comovement ($k = 0$) is when economic and floating population changes occur simultaneously. Finally, antecedence ($k > 0$) refers to when floating population changes occur following economic changes. The cross-correlation analysis results were obtained regarding precedence and comovement, excluding antecedence cases. When performing a cross-correlation analysis between commercial transit floating population and lodging and restaurant indicators, the value of k was found to be -2, implying that there is precedence wherein the floating population changes before the development composite indices. In other words, this implies that lodging and restaurant indicators have changed according to commercial transit floating population changes over the last two months.

Table 6 presents a precedence wherein the indicators of service industries related to lodging, restaurants, specialty retailers, arts, sports, and leisure change following business and shopping-related floating populations. It demonstrates the potential of this study, whose purpose was the creation of precedent prompt development composite indices using real-time data.

In contrast to correlation coefficients, cross-correlation coefficients show the numerical value of the degree of morphological similarity between two variables. Therefore, correlation and cross-correlation coefficients exhibit different values; thus, a high cross-correlation coefficient should not be interpreted the same as a high correlation coefficient. For example, commercial floating population and specialty retail indicators exhibit a high correlation coefficient (0.81), implying that they significantly influence each other's changes. However, their cross-correlation coefficient (0.78) and LAG (−2) imply that commercial floating population changes increase lodging and restaurant indicator values after two months and have a very similar pattern of 0.78. In addition, there is also the possibility of obtaining a very high and relatively low correlation coefficient of 0.90 and 0.50, respectively. Therefore, the purpose of this study was to obtain different results by selecting relatively high indicators wherein both variables have a relatively high average of 0.75 or greater. Table 7 presents nine indicators according to the floating population categories selected in this study: commercial floating population and lodging and restaurant and specialty retail service industries; short distance commercial floating population and specialty retailers; shopping floating population and lodging and restaurants, and art/sport and leisure-related industries; short distance shopping floating population and transportation and warehousing, lodging and restaurants, and art/sport and leisure industries; and long-distance shopping floating population and lodging and restaurant businesses.

Although cross-correlation coefficients exhibit a difference based on each indicator, the absolute value of correlation coefficients was found to be relatively higher than other variables in the case of shopping, short-distance shopping floating population, and lodging and restaurant businesses. In contrast, commercial, lodging, and restaurant businesses, as well as long-distance shopping-related lodging and restaurant businesses, exhibited relatively low values.

In the cross-correlation analysis, most floating population types and their development composite indices exhibited a positive (+) correlation with a short lag of 2 months or less. It implies that development composite indices change proportionally to floating population changes within two months. Consequently, this indicates that economic revitalization and recessions are directly connected to floating population trends. However, prompt and appropriate policies such as the alleviation of e-commerce policies when national disasters such as COVID-19 occur must be prepared as they reduce the amount of floating population, which is directly connected to the occurrence of economic recessions.

4.2. Multiple Regression Model Results. A multiple regression model was created to analyze and determine whether the floating population and corona cases per day affect the development composite index changes, and if they do, then to which degree. Results of regression analysis on the indicators selected for analysis shown in Table 8 showed that a floating population variable in the five composite indices models was significant to the confidence level. The five composite economic index models are as follows:

Table 6: Result for cross-correlation analysis.

Lag	Commercial		Shopping	
	Accommodation and restaurant business	Professional retail store	Accommodation and restaurant business	Arts, sports, and leisure services
−5	0.1059	−0.0219	0.1944	0.2612
...	...	...	...	...
−2	0.5822	0.7844	0.4911	0.6893
−1	0.5412	0.5145	0.4353	0.7173
0	0.4792	0.1914	0.3154	0.7170
1	0.2966	−0.0657	0.1835	0.5891
2	0.2777	0.0234	0.1639	0.4849
...	...	...	...	...
5	0.1509	0.2924	0.0819	0.0244

(1) Lodging and restaurants for shopping purposes

(2) Art/sport and leisure-related businesses for shopping purposes

(3) Short-distance transportation and warehousing for shopping purposes

(4) Short-distance lodging and restaurants for shopping purposes

(5) Short distance art/sport and leisure-related businesses for shopping purposes

In addition, the floating population regression coefficients were all positive, implying that development composite indices increase proportionally with the floating population. Therefore, it is interpreted as development composite index values increasing by a factor of 0.319 each time the floating population value of short distance art/sport and leisure-related services for shopping purposes increases by 1. In contrast, none of the COVID-19 coefficients were found to be significant to a confidence level of 95% in the models. It is because the number of Corona cases per day in Ulsan Metropolitan City is remarkably low than the number of confirmed cases nationwide, thereby preventing it from significantly affecting the floating population value; also, the sense of crisis among people concerning the pandemic dwindles which generates limitations for it to be reflected statistically.

Based on the results presented in Table 8, the actual value (Y) of development indicators concerning floating populations with shopping and short-distance shopping transit purposes and the value ($\hat{Y}$) predicted through the regression model are shown in Figure 2.

Quantitative indicators such as mean average error (MAE), mean square error (MSE), root mean square error (RMSE), and root mean square percentage error (RMSPE) were used to confirm the fit of the models. Each indicator is defined as follows: $1/n\sum_{i=1}^{n}|y_i^{\text{real}} - y_i^{\text{pred}}|$ for MAE, $1/n\sum_{i=1}^{n}(y_i^{\text{real}} - y_i^{\text{pred}})^2$ for MSE, $\sqrt{\text{MSE}}$ for RMSE, and $\sqrt{1/n\sum_{i=1}^{n}(y_i^{\text{real}} - y_i^{\text{pred}}/y_i^{\text{real}})^2}$ for RMSPE, where n is the number of samples, y_i^{real} is an actual composite economic index, and y_i^{pred} is a composite economic index predicted by the model of this study.

Although the predicted value is generally similar to the actual value, the monthly error value appears rather large, thereby rendering the prediction model used challenging to be considered optimal. However, among all the predicted models, the MSE value presented in Table 9 for lodging and restaurant businesses for shopping purposes is 49.03, and that of the model for short-distance shopping purposes is 45.82, which shows a relatively low error in comparison with other indicators. Therefore, it can be concluded that lodging and restaurant indicators obtained from regression analysis results were relatively appropriate to predict economic situations considering the floating population with shopping purposes.

4.3. Time Series Analysis Results. Floating population and development composite index data are defined as time series owing to data observed over time. However, monthly and quarterly floating population data have regular seasonal factors (depending on the season, climate, and holiday). Thus, a seasonal adjustment process is necessary to perform an accurate travel pattern analysis. Figure 3 shows that the stationarity of time series data with seasonal variations was secured by diagnosing the seasonality of floating population data and evaluating the disparity between seasons before analyzing the correlation with development composite indices to identify the trend and cyclical changes.

The regression analysis model found that the number of daily COVID-19 cases, which is an exogenous variable, is not significant to the previously specified confidence level. The SARIMAX model, which allows the prediction of indicators when exogenous variables with time differences exist, was used to compensate for this limitation. Considering that the data used in this research is the monthly data collected for two years with a sample number lower than 30, its statistical significance was examined, and the confidence interval was adjusted from 95 to 80%. All d values, one of the SARIMAX model values (p, d, q, P, D, and Q) that represent the order of difference, were set to 1, because data stationarity was already secured through one differentiation. Furthermore, as the interval of time series (s) refers to the monthly data, it was set to 12. Consequently, following model establishment, the t value of indicators and exogenous variables was found

Table 7: Summary of correlation and cross-correlation results.

Market division	Composite index	Division	Lag	Correlation value	Cross-correlation value	Mean
Commercial	Accommodation and restaurant business	Leading	−2	0.93	0.58	0.75
	Professional retail store	Leading	−2	0.81	0.78	0.79
Commercial with short distance	Professional retail store	Leading	−2	0.83	0.79	0.80
Shopping	Accommodation and restaurant business	Leading	−2	0.94	0.73	0.84
	Arts, sports, and leisure services	Leading	−1	0.88	0.71	0.79
Shopping with short distance	Transportation and warehousing industries	Leading	−1	0.85	0.73	0.79
	Accommodation and restaurant business	Leading	−1	0.94	0.73	0.84
	Arts, sports, and leisure services	Accompanying	0	0.88	0.72	0.80
Shopping with long distance	Accommodation and restaurant business	Leading	−2	0.90	0.59	0.75

Table 8: Regression model of composite economic indexes.

Market division	Composite index	Floating population coefficient	P value
Shopping	Accommodation and restaurant business	0.203	0.001
	Arts, sports, and leisure services	0.271	≤0.001
Shopping with short distance	Transportation and warehousing industries	0.116	0.001
	Accommodation and restaurant business	0.245	≤0.001
	Arts, sports, and leisure services	0.319	0.001

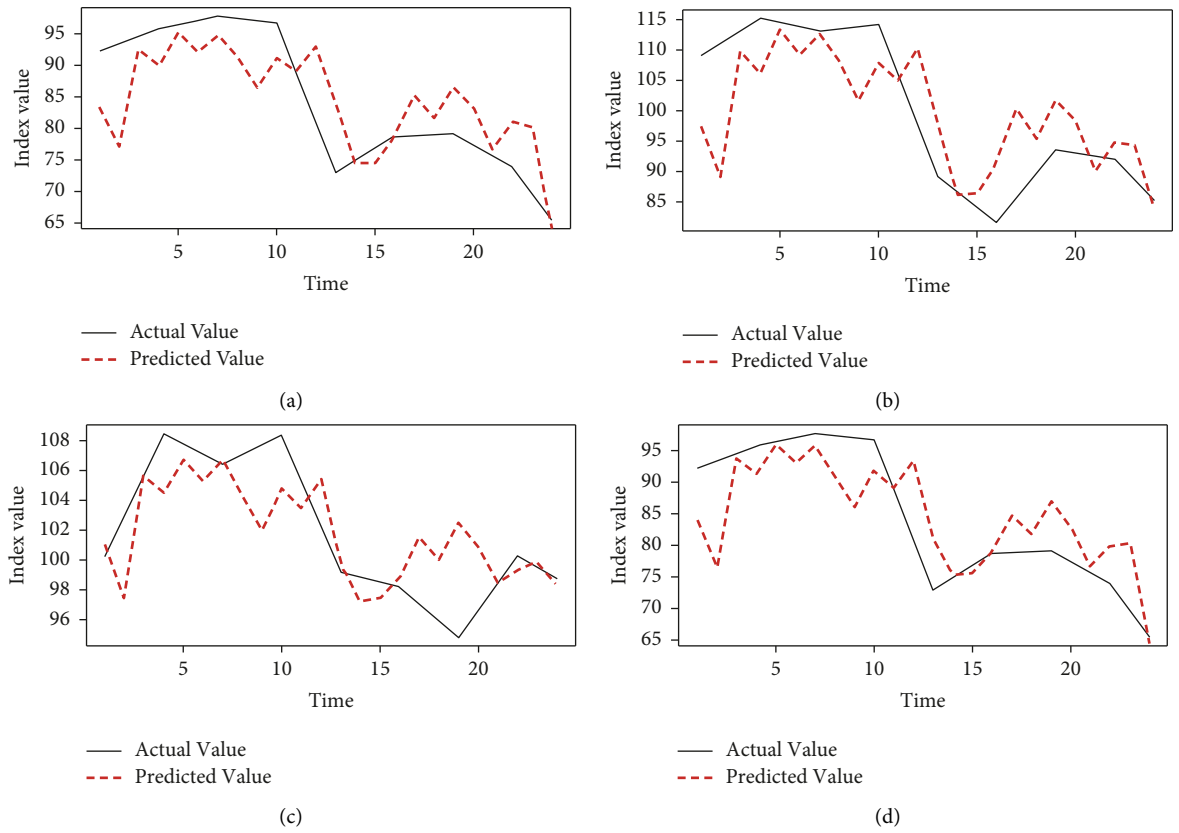

Figure 2: Comparison between actual and predicted values for regression models. (a) Shopping accommodation and restaurant business. (b) Shopping arts, sports, and leisure services. (c) Shopping with short distance transportation and warehousing industries. (d) Shopping with short distance accommodation and restaurant business.

Table 9: Accuracy tests for regression models.

Market division	Composite index	MSE	RMSE	RMSPE	MAE
Shopping	Accommodation and restaurant business	49.03	7.00	0.12	5.50
	Arts, sports, and leisure services	284.44	15.54	0.26	16.87
Shopping short	Transportation and warehousing industries	368.83	17.25	0.28	19.20
	Accommodation and restaurant business	45.82	5.13	0.10	6.77
	Arts, sports, and leisure services	280.75	15.60	0.25	16.76

to be significant to the confidence interval for the remaining p and q values, and AIC and BIC values were set to be the lowest values.

Similar to the regression analysis results, SARIMAX model analysis also showed the negligible influence of the COVID-19 exogenous variable. Thus, another analysis was

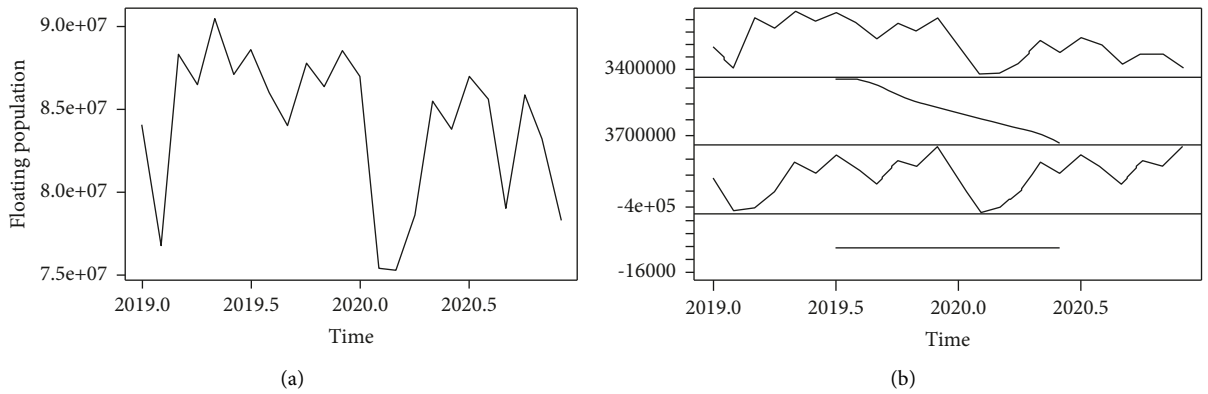

Figure 3: Time series graphs with random, seasonal, and trend components in the floating population. (a) Floating population trend. (b) Seasonal diagnosis.

Table 10: Result of time series analysis.

Market division	Composite index	(P, D, Q)	(p, d, q)	Floating population's coefficient	MA's t value
Commercial	Accommodation and restaurant business	(0, 0, 1)	(0, 1, 0)	0.0237	1.65
	Professional retail store	(0, 0, 1)	(0, 1, 0)	0.0082	4.53
Commercial with short distance	Professional retail store	(0, 0, 1)	(0, 1, 0)	0.0104	4.29
Shopping	Accommodation and restaurant business	(0, 0, 2)	(0, 1, 0)	0.0766	5.03/3.0
	Arts, sports, and leisure services	(0, 0, 2)	(0, 1, 0)	0.0943	4.21/2.80

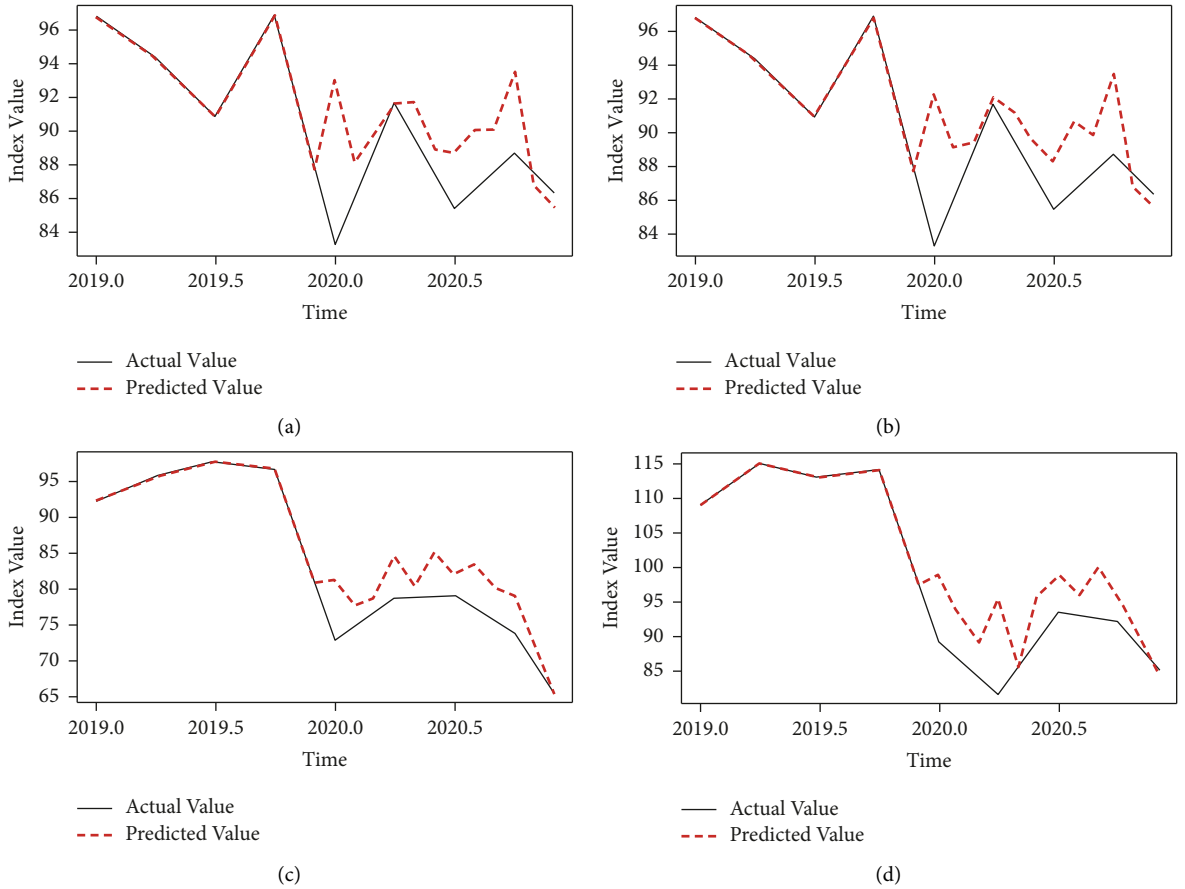

Figure 4: SARIMAX's results of actual value and predicted value. (a) Commercial professional retail store. (b) Commercial short professional retail store. (c) Shopping accommodation and restaurant business. (d) Shopping arts, sports, and leisure services.

Table 11: Accuracy tests for SARIMAX model.

Market division	Composite index	Division	MSE	RMSE	RMSPE	MAE
Commercial	Accommodation and restaurant business	Leading	39.19	6.26	0.11	3.91
	Professional retail store	Leading	6.76	2.6	0.04	1.38
Commercial with short distance	Professional retail store	Leading	6.35	2.52	0.04	1.37
Shopping	Accommodation and restaurant business	Leading	10.63	3.26	0.06	2.02
	Arts, sports, and leisure services	Leading	21.16	4.60	0.10	2.67

performed by excluding COVID-19 and adding the floating population as the exogenous variable of the SARIMAX model. The SARIMAX (P, D, Q) (p, d, q) [12] model results are presented in Table 10.

The confirmed SARIMAX model was created using floating population as exogenous variable and development composite index values, and monthly development composite indicator figures were verified for 24 months, from January 2019 to December 2020.

Figure 4 shows the actual value (Y) of each development indicator of floating populations related to commercial, shopping, and short-distance commercial purposes and the predicted values $(\hat{Y})$ obtained through the time series model. The indicator values predicted using the time series model were found to be approximately equal to the actual values of 2019; however, an observable difference was found during the spread of COVID-19.

Nevertheless, as shown in Table 11, among time series models, the maximum MSE value was 40, which is lower than its minimum value of 45 that was obtained through regression analysis. Thus, it is evident that the time series model exhibits a relatively better performance than the regression model in predicting future development indices. Moreover, the time series prediction model found that commercial and short-distance commercial floating population is the best suited for explaining specialty retail service indicators. Thus, it can be concluded that as a precedence indicator, it is valid for use as a prompt development composite indicator to quickly and accurately predict future situations.

5. Conclusion

Development composite indices are created by processing and synthesizing indicators that reflect economic conditions well, such as employment, productivity, consumption, investment, foreign affairs, and finance, and they can quickly identify national economic trends. However, the time of publication of existing development composite indices is concentrated at the end of each month, quarter, and year, which acts as a limitation in terms of identifying quickly shrunken regional economic situations (such as the prolongation of COVID-19 and the deterioration of the working-class economy) and accordingly establish prompt economic policies and corporate countermeasures. To compensate for these limitations, research on developing indicators that can diagnose regional economic changes quicker using a variety of real-time traffic big data is ongoing.

In this study, floating population data of Ulsan Metropolitan City in South Korea from 2019 to 2020 was used to estimate prompt development composite indices. Various statistical methods, such as correlation, regression, and time series analyses, were used to define the development composite index correlated to real-time floating population data and the type of floating population with high predictability of economic situations. Consequently, prompt development composite indices and models appropriate to predict economic situations were determined.

Indicators with high average correlation and cross-correlation values were primarily selected to develop an economy prediction model. Subsequently, regression analysis revealed that economic indicators could be predicted when combining floating population data with transportation, warehousing, lodging, restaurant, art/sports, and leisure industry data. Furthermore, through time series analysis results, lodging, restaurant, art/sports, leisure, and specialty retailer data were obtained, which is expected to secure higher reliability for regional development index prediction when combined with floating population data. In addition, a comparison of MSE revealed the time series model's prediction accuracy to be higher than that of the regression model. Thus, based on these results, the floating population was used to classify economic situation prediction indices into three types: (1) lodging and restaurants, (2) specialty retailers, and (3) art/sport and leisure businesses.

The finally selected development composite indices were all precedent indicators wherein the floating population changed before regional development indices. Thus, they can be used to diagnose time series regional economic trends and are expected to be used as standards to predict reliably future economic situations. However, the floating population data based on communication data used in this study has several limitations, such as the low number of samples owing to the short two years (2019 to 2020). In addition, although the spread of COVID-19 had a significant socioeconomic effect, its influence could not be used as an exogenous variable in the model because the amount of daily COVID-19 cases in Ulsan Metropolitan City was too low, which, in time, changed the perception of people on the level of risk of epidemics. Therefore, in future research, collecting and analyzing samples for a more extended period is necessary, predicting regional composite indices using advanced model methods and considering various exogenous factors that affect socioeconomics.

Data Availability

The data that support the findings of this study are available from Korea Transport Institute, but restrictions apply to the

availability of these data, which were used under license for the current study, and so are not publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Dongho Kim and Jungyeol Hong were responsible for research conception and design. Jieun Na and Youjeong Kang were responsible for data collection. Jungyeol Hong, Jieun Na, and Youjeong Kang were responsible for analysis and discussion of results. Jungyeol Hong and Jieun Na were responsible for draft manuscript writing. All authors approved the final version of the manuscript.

Acknowledgments

The study was funded by the Korea Transport Institute.

References

[1] H. S. Park, S. H. Cho, N. M. Hong, and B. H. Jang, *A Study on the Development of the Seoul Metropolitan Government Composite Index*, Seoul Research Institute, Seoul, South Korea, 2006.

[2] S. Šćepanović, I. Mishkovski, P. Hui, J. K. Nurminen, and A. Ylä-Jääski, "Mobile phone call data as a regional socio-economic proxy indicator," *PLoS One*, vol. 10, no. 4, Article ID e0124160, 2015.

[3] I. Arhipova, G. Berzins, E. Brekis, J. Binde, and M. Opmanis, "Mobile phone data statistics as proxy indicator for regional economic activity assessment," *FEMIB*, vol. 37, pp. 27–36, 2019.

[4] I. Arhipova, G. Berzins, E. Brekis et al., "Mobile phone data statistics as a dynamic proxy indicator in assessing regional economic activity and human commuting patterns," *Expert Systems*, vol. 37, no. 5, Article ID e12530, 2020.

[5] I. Arhipova, G. Berzins, A. Erglis, E. Ansonska, and J. Binde, "Socio-economic situation in Latvia's municipalities in the context of administrative-territorial division and unexpected impact of COVID-19," *Journal of Global Information Management*, vol. 30, no. 10, pp. 1–27, 2022.

[6] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti, *Estimating Economic Development with mobile Phone Data*, 2016,http://new.cisstat.org/web/guest.

[7] J. E. Blumenstock, "Estimating economic characteristics with phone data," *AEA papers and proceedings*, vol. 108, pp. 72–76, 2018.

[8] G. Kreindler and Y. Miyauchi, "Measuring commuting and economic activity inside cities with cell phone record," *The Review of Economics and Statistics*, vol. 105, pp. 1–48, 2021.

[9] L. Dong, S. Chen, Y. Cheng, Z. Wu, C. Li, and H. Wu, "Measuring economic activity in China with mobile big data," *EPJ Data Science*, vol. 6, pp. 29–17, 2017.

[10] F. Van Ruth, *Traffic Intensity as Indicator of Regional Economic Activity*, Statistics Netherlands, The Netherlands, 2014.

[11] M. S. Arslanalp, M. M. Marini, and M. P. Tumbarello, *Big Data on Vessel Traffic: Nowcasting Trade Flows in Real Time*, International Monetary Fund, Washington, DC. USA, 2019.

[12] E. Rowland, "Faster indicators of UK economic activity: road traffic data for England," *Data Science Campus Report*, Office for National Statistics, 2019.

[13] C. P. Dancey and J. Reidy, *Statistics without Maths for Psychology*, Pearson/Prentice Hall, Hoboken, NJ, USA, 2007.

[14] J. Y. Ham and J. Y. Son, "Causality between housing price and policy: is housing policy exogenous?" *Housing Policy Studies*, vol. 20, no. 4, 2012.

[15] J. Hong, E. Han, C. Choi, M. Lee, and D. Park, "Estimation of shared bicycle demand using the SARIMAX model: focusing on the COVID-19 impact of seoul," *The Journal of The Korea Institute of Intelligent Transport Systems*, vol. 20, no. 1, pp. 10–21, 2021.