Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 7383 - 7401

Research Article

An Analysis of Public Transportation Satisfaction for Fine Dust Reduction

Seung-Yeon Hwang[a], Jeong-Joon Kim[b]

[a] Dept of Computer Engineering, University of Anyang, Anyang-si, Gyeonggi-do, Republic of Korea
[b] Corresponding Author, Dept of Software, University of Anyang, Anyang-si, Gyeonggi-do, Republic of Korea

Abstract

Korea has focused its efforts on reducing fine dust particles in Seoul over the past day, with the focus on improving the quality of fine dust products, such as a complete reduction in public transportation, closing down power plants, and exchanging fine dust with China. Despite these efforts, however, the nation's fine dust concentration has not significantly decreased, leaving people feeling uneasy and dissatisfied. Therefore, this paper utilizes a multiple regression analysis conducted by using R, a tool for analyzing big data, to identify variables that affect the level of public transport satisfaction, and to analyze multiple regression analysis at each level.

Keywords: Big Data, R, Multiple regression, Public Transportation, Dust, Reduction

1. Introduction

As the problem with fine dust concentration has become more serious recently, the government has set a goal of reducing fine dust by 30 percent by 2022. To this end, the government has set up measures to provide free public transportation in Seoul, restrict the operation of old diesel cars, expand the two-part vehicle system, lower the height of air pollution monitoring stations, increase the number, and provide subsidies for eco-friendly boilers. In China, which thought it was the cause of fine dust, the joint fine dust reduction project between South Korea and China finally began by expressing their willingness to reduce fine dust. However, air pollution in Korea is still at a serious level, and public complaints and inconveniences are increasing. Moreover, cities such as around the industrial complex have more serious damage. As a result, according to the results of the "Achievement Survey by 100 State Tasks" conducted by the Korea Public Policy Evaluation Association, experts had mixed opinions on the policy of people's livelihoods that the president was most interested in. The third national goal, "the country responsible for my life," received low scores in the case of state affairs, such as fine dust measures and strengthening the public nature of education, while medical publicness and others scored generous points. According to a detailed evaluation of the top 100 tasks, the 58th national task, "Creating a pleasant atmospheric environment without worrying about fine dust," was 73.75 points, the lowest among the 100 national tasks.

Most of the fine dust comes from China, but the main factors of fine dust in Korea are factories, power plants, car exhaust gases, construction sites, and incineration plants that burn fossil fuels such as coal and oil[1].

In response to the risk of air pollution, Korea also revised the Air Environment Conservation Act to manage fine dust during air pollution, and introduced ultrafine dust forecasts in eight cities and provinces, including Seoul and Gyeonggi Province[2].

On January 15, 2018, the Seoul Metropolitan Government introduced an unconventional policy of free public transportation during the office-going hour. Due to the "free public transportation" policy, the number of public transportation users has increased, while the number of passenger cars has decreased. According to the Seoul Metropolitan Government's analysis of traffic at 14 locations in Seoul, the number of vehicles entering the city decreased by 2,099 (1.8%) compared to the same day last week and reduced 0.8 tons of fine dust.

Therefore, in this paper, we analyze the satisfaction of public transportation in order to expect the long-term effect of reducing traffic rather than expecting short-term effects such as free public transport policies. Through the analysis results, the company aims to reduce traffic volume and eventually reduce fine dust by increasing the utilization rate of public transportation.

First of all, after analyzing the overall public transportation usage rate in time series, the bus usage rate was increasing. Therefore, we hope to find and analyze factors that increase subway utilization, one of the largest means of public transportation, and expect greater effects. On the contrary to buses, subway utilization was decreasing when looking at time series data and finding factors that could increase the utilization of subways that do not use fossil fuels would have a more positive effect in the long run. Data quantifying factors such as fares, amenities, congestion, and air quality, which are closely related to utilization, were collected from the Seoul Public Data Portal and analyzed how each factor correlates with utilization through multiple linear regression. And I would like to present policies on long-term effects.

II. RELATED TECHNOLOGIES

1. Big Data

The current data is doubling the amount every year in zettabytes. Big data involves a technique to store a lot of this data and analyze it with various big data tools to extract meaningful content. The steps dealing with big data can be largely divided into the steps of collection, storage, processing, analysis, and visualization[4].

2. Distributed File System

Distributed file systems are physically connected to different computer networks, allowing users to use multiple file systems as a single file system. Furthermore, with advances in network technology and active dissemination, several small computers are networked together, making it easy to share information between users and efficiently using storage space without time and place constraints[5].

3. Map Reduce

MapReduce was released in 2004 as a software framework for parallel/distributed processing of big data and is represented by Hadoop. It is now widely used by companies dealing with big data such as Amazon, Yahoo, and Facebook. A map step is a step that transforms scattered unstructured data into a key/value pair in a formalized form by tying them together with related data. In the Reduce phase, the task is to extract useful data, such as removing redundant data, by performing parallelism on key/value pairs in a structured form created in the map phase. Basically, MapReduce is effective in handling big data by performing batch-based processing[6].

4. Statistical Language R

Although R is program originally used as a statistical tool, it is easier to use than languages such as C and JAVA, and because it is open-source, other user-created packages can be easily used. Furthermore, we can implement state-of-the-art data mining techniques with packages that provide visual analysis results. R is a language adopted by companies that seek high-performance big data analysis such as IBM and Teradata. R can easily apply data-appropriate analytical techniques by downloading packages using the Internet[7].

III. RELATED RESEARCH

1. Analysis of factors affecting user satisfaction of public transportation[8].

Existing literature conducted only basic analysis of public transportation satisfaction, but the factor analysis study on public transportation user satisfaction was studied based on the socioeconomic characteristics of users.

The study aims to compare and analyze the distribution of satisfaction and identify socioeconomic factors on satisfaction based on the Ministry of Land, Transport and Maritime Affairs' recently conducted 2011 Transportation User Satisfaction Survey Data on the Seoul Metropolitan Area. To this end, various descriptive statistical analyses were carried out using data such as satisfaction survey results and respondents' use of public transportation and socio-economic characteristics, and an ordered probit model was established for satisfaction. As a result, it showed that the characteristics of respondents such as gender, age, income, occupation, and educational background, as well as the region, main means of transportation, transportation card use, and passage routes affect user satisfaction. The advantages are that various factors can be incorporated into correlation analysis and that there are many factors related to public transportation satisfaction around. And based on that, we present policy directions. However, since the data is old and limited to a single year of 2011, only research on the situation at that time is possible. However, this paper uses various data from 2005 to 2016 to confirm changes and uses regression analysis to examine the relationship between convenience facilities, fine dust, carbon dioxide, formaldehyde, and carbon monoxide.

2. A study on the countermeasures against fine dust using the foundation of big data utilization[9].

The study was based on a pilot project of big data flagship that lasted for about four months from August 30, 2017 to December 20, 2017.

Research and development aims to build a foundation for fine dust response by designing the basis for public safety and behavior from fine dust through pilot operation of roadsides and building nests in data utilization institutions (Jeju-do Office, Changwon City Hall, and Gwangmyeong City Hall). In addition, statistical analysis using 'Deep Learning' was utilized to derive countermeasures against fine dust according to the purpose of Big Data utilization foundation. Through the analysis of deep learning statistics, we present countermeasures against the analysis of fine dust vulnerable zones and the spread of fine dust. In addition, by supplementing the existing national measurement network, we intend to provide useful services to the people by complementing the measured air quality data and national weather data based on IoT.

This study can be noted that it is a countermeasure service in the event of fine dust through various data analysis. Such a service must be a very good study in that people do not have to find out how to deal with fine dust themselves. However, if the fine dust level continues to soar and there is a situation where people have to wear masks every day, they should also consider what to do. Therefore, this paper deals with the consideration of why public transportation utilization is decreasing and how public transportation utilization can be increased, which is a fundamental rising factor of fine dust.

IV. Analysis

The purpose of this paper is to analyze how various factors correlate to public transport satisfaction. Multi-linear regression analysis is needed to achieve the objective of the study. Multi-linear regression is the creation of a regression model for predicting dependent variables with multiple independent variables, and basically the dependent variable must be a continuous variable.

Table 1. Statistical Data of Transportation for Commuting

Data structure
'data.frame':   12 obs. of 10 variables: $ Period: int 2005 2006 2007 2008 2009 2010 2011 ... $ Walking: num 20.6 18.3 17.6 19 19.1 16.8 13.2 15.3 ... $ Bicycle: num 1.8 2 1.3 2.3 2.6 2.7 1.9 3.1 2.7 2.8 ...

```
$ Motorcycle: num  0 0 0.9 1 0.8 0.6 0.6 0.3 0.6 0.7 ...
$ Bus: num  19.7 20.7 20.6 19.6 19 22.5 16.8 20.9 ...
$ Subway: num  15.4 19.1 18 15.1 14 12.8 12.8 14.6 ...
$ Bus+Subway: num  17.7 14.8 15.8 19.1 18.6 20.6 25 ...
$ Taxi: num  0.5 0.5 0.7 0.3 0.4 0.4 0.2 0.1 0.3 0.2 ...
$ Car: num  20.4 21.2 20.9 20 22.6 21.3 26 21.1 ...
$ Etc: num  0.2 0.3 0.4 0.4 0.3 0.2 0.2 0 0 0 ...
```

The period variables in Table 1 refer to the year 2005 to 2016, walking, cycling, motorcycle, bus, subway, taxi, passenger car, and other variables refer to the rate of use of each chatter used for commuting and commuting.. Bus and subway variables refer to the utilization rate of both buses and subways for commuting and commuting, and the data provided by the Seoul Open Data Square Portal are refined as shown in the table above. The reason for using data during commuting and commuting hours is that you can check which transportation you use the most during the time when the population moves the most.

Table 2. Execution result source code

Execution result source code

```
Transportation<-read.csv("transportation.csv", sep=",", header=T)
Transportation_seoul<- Transportation[Transportation$Classification=="Seoul City",]
Transportation_seoul<- Transportation_seoul[,-c(2,3,6,13,14)]
Transportation_seoul<- Transportation_seoul[c(order(Transportation_seoul$Period)),]
install.packages("ggplot2")
library("ggplot2")
ggplot(Transportation_seoul, aes(x=Period, y=Bus)) +
xlab('YEAR') +
ylab('Rate') +
geom_line(colour="red") +
geom_point(size=2, shape=19, colour="red") +
theme_bw() +
scale_x_continuous(breaks =c(2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016)) +
ggtitle("Annual bus utilization rate") +
theme(plot.title = element_text(hjust = 0.5))
```

ggplot2 Using the ggplot function provided in the library, we visualize the distribution of transportation statistics used to commute and commute to Seoul as shown in Figure 1. Figure 1 shows that bus usage is increasing while subway usage is decreasing. The purpose of this study is to reduce the high utilization of private cars and increase the utilization of public transportation such as buses and subways.

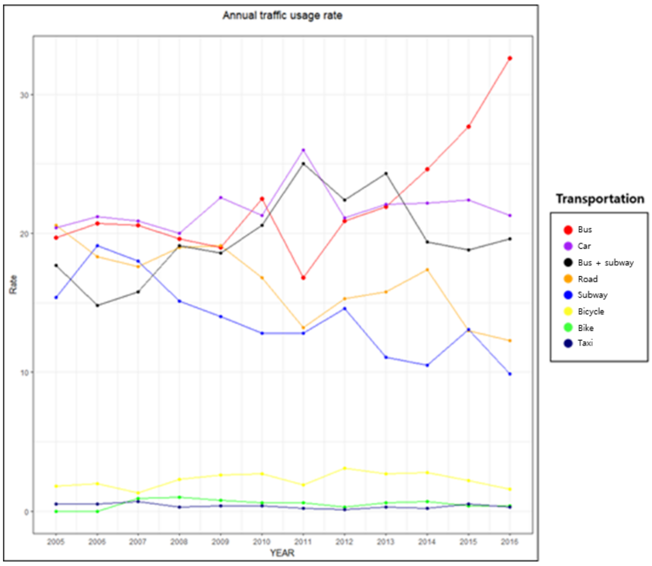

Fig. 1. ggplot Visualization

First, in order to check whether the increase in bus utilization and bus satisfaction were related, the Seoul Metropolitan Government's bus use satisfaction statistics data provided by the Seoul Open Data Square were refined and used.

Table 3. Data on the satisfaction level of buses in Seoul

Period	**Synthesize**	**Transfer Use**	**Station Location**	**Bank Time**	**Time of arrival**	**Fare**
2011	5.51	6.67	6.10	6.02	5.63	5.18
2012	6.03	6.82	6.14	6.13	5.72	5.33
2013	5010	6.91	6.15	6.20	5.77	5.43
2014	5.17	7.03	6.16	6.27	5.86	5.53
2015	5.23	7.13	6.17	6.33	5.96	5.62
2016	5.35	7.31	6.20	6.43	6.09	5.77

The period variables in Table 3 are for the years 2011 to 2016, and the overall variables represent the mean, transfer use, stop location, operating time, and arrival time of the factors affecting each satisfaction. A fare variable refers to a score that expresses satisfaction with each factor out of 10.

Table 4. Execution result source code

Execution result source code

```
Satisfaction<-read.csv("satisfaction.csv", sep=",", header=T)
Satisfaction_seoul<- Satisfaction[Satisfaction$Classification=="Seoul City",]
Satisfaction_seoul<- Satisfaction_seoul[,-c(2,3)]
Satisfaction_seoul<- Satisfaction_seoul[c(order(Satisfaction_seoul$Period)),]
Transportation_seoul_comparison<- Transportation_seoul[7:12,c(1,5)]
Satisfaction_seoul_comparison<- Satisfaction_seoul[,c(1,2)]
Transportation_seoul_comparison<- transform(Transportation_seoul_comparison,
        Bus = scale(Bus))
Satisfaction_seoul_comparison<- transform(Satisfaction_seoul_comparison,
        Synthesize = scale(Synthesize))
ggplot(Transportation_seoul_comparison, aes(x=Period, y=Bus)) +
xlab('YEAR') +
ylab('Rate') +
geom_line(colour="red") +
geom_line(aes(x=Satisfaction_seoul_comparison$Period,y=Satisfaction_seoul_comparison$Syntheisize),colour="Orange") +
geom_point(size=2, shape=19, colour="red") +
geom_point(aes(x=Satisfaction_seoul_comparison$Period,y=Satisfaction_seoul_comparison$Synthesize),colour="Orange") +
theme_bw() +
scale_x_continuous(breaks = c(2011,2012,2013,2014,2015,2016))+
ggtitle("Utilization according to annual bus satisfaction") +
theme(plot.title = element_text(hjust = 0.5))
```

Using the ggplot function, we can see how bus utilization varies with bus satisfaction as shown in Figure 2. Bus variables, which mean bus utilization of transportation statistical data used for commuting and commuting, and comprehensive variables, which mean average satisfaction of bus use of Seoul bus satisfaction statistical data, were visualized after standard regularization. Figure 2 shows that bus satisfaction changes and bus utilization changes have a positive relationship.

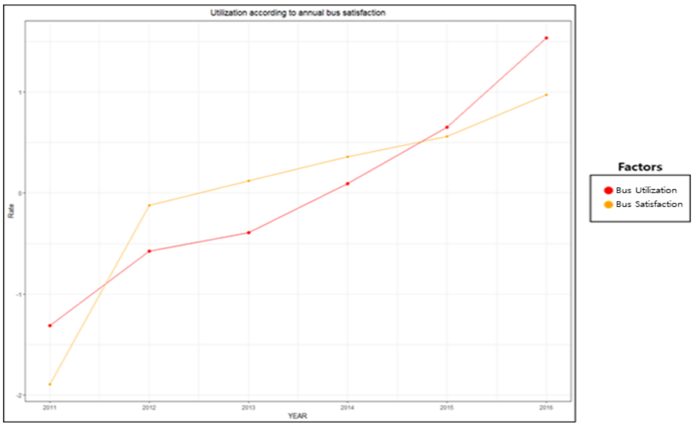

Fig. 2. ggplot Visualization

Figure 3 visualizes each of the factors in Seoul's bus use satisfaction statistical data. Through the graph, bus users showed high satisfaction in terms of transfer use. In addition, it is not yet high satisfaction, but it can be seen that satisfaction increases every year in terms of driving time, arrival time, and fare. On the other hand, the location of the bus stop is interpreted as maintaining parallel lines, so if the location of the bus stop is adjusted to reflect the opinions of the users, the overall bus satisfaction increases and thus the bus utilization rate increases.

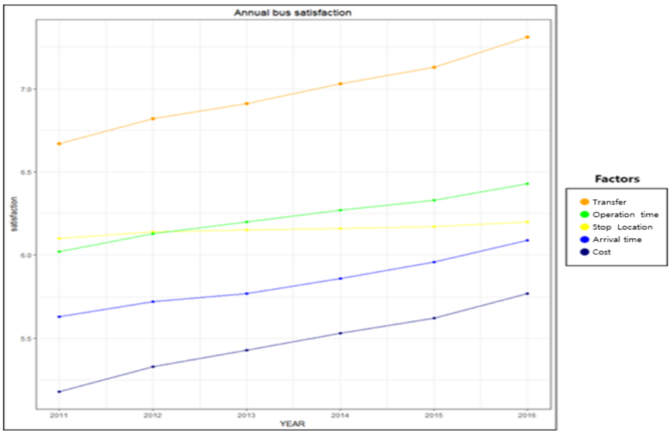

Fig. 3. ggplot Visualization

While bus utilization is increasing, only the variables associated with buses and subways are extracted from the transportation statistical data used to commute and commute to find factors affecting the decreasing subway utilization, and visualized as shown in Figure 4.

In Figure 4, bus+subway utilization shows a similar pattern of decline to subway utilization since 2011. While bus and subway usage rates are decreasing, it can be interpreted that people who used to use the bus and subway together gradually do not use the subway, but only buses.

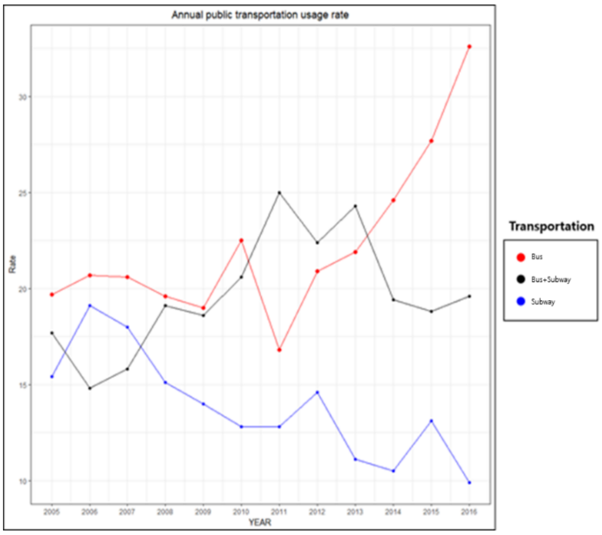

Fig. 4. ggplot Visualization

Before analyzing the factors affecting subway utilization, relevant public data were used to check whether subway satisfaction and utilization were related.

Table5. Current data on transportation personnel by year in Seoul

Year	Line 1	Line 2	Line 3	Line 4	Line 5
2005	171943936	698775955	259165710	306528741	359103586
2006	169903083	699222041	257543491	304327475	357749023
2007	169635080	707328238	256172707	298621431	357939364
2008	168096727	727057819	256547773	295222107	361731107
2009	163860092	732037588	257501239	297131821	362632685
2010	164409302	731847885	275466721	303625029	368837234
2011	170110521	747577667	282997903	308842571	377382166
2012	167784425	752917456	286036938	306926434	378416313
2013	168206476	759307998	289692822	306485363	380923165
2014	170807317	771241619	293775202	308201207	386006336
2015	164138852	761807002	286360778	302565175	378717952
2016	164634048	762447319	287363289	303813353	379564502

The year variables in Table 5 are the annual number of people transported by each subway line in Seoul for the years 2005 to 2016, 'Year', 'Line 1', 'Line 2', and 'Line 4', and the average variable is the average annual number of people transported by the Seoul subway. The data on the number of people transported by year in Seoul provided by the Seoul Open Data Plaza has been refined as shown in the table above.

Table 6. Statistical Data on Traffic Satisfaction in Seoul

Period	Public Transport Satisfaction	Bus	Subway	Taxi
2005	5.52	5.61	6.26	4.7
2006	5.6	5.66	6.3	4.83
2007	5.75	6	6.33	4.92
2008	5.83	6	6.3	5.14
2009	6	6.19	6.52	5.27
2010	6.19	6.16	6.71	5.67
2011	6.23	6.28	6.79	5.61
2012	6.31	6.58	6.81	5.54
2013	6.4	6.67	7.03	5.5
2014	6.47	6.78	6.95	5.7
2015	6.59	6.88	7.01	5.88
2016	6.39	6.86	6.71	5.6

The period variable in Table 6 is the year from 2005 to 2016, the public transportation use satisfaction variable means the average satisfaction level of public transportation such as bus, subway, and taxi, and the bus, subway, and taxi variables mean each satisfaction score expressed on a scale of 10 points. In addition, the statistical data on traffic use satisfaction in Seoul provided by the Seoul Open Data Plaza were refined as shown in the table above.

Table 7. Execution result source code

Execution result source code

```
Transportation_personnel<- read.csv("Transportation_personnel.csv", sep=",", header=T)

Transportation_personnel<- Transportation_personnel[complete.cases(Transportation_personnel),]

Transportation_personnel<- Transportation_personnel[Transportation_personnel$ division=="Number of people transported(person)",]

Transportation_personnel<- Transportation_personnel[c(order(Transportation_personnel$ Year)),]

Transportation_personnel<- Transportation_personnel[c(21:32),c(1,3,4,5,6)]

install.packages("magrittr")

install.packages("dplyr")
```

```
library(magrittr)

library(dplyr)

Transport_Average<- Transportation_personnel %>% mutate(Average = (XLine 1 + XLine 2 + XLine 3 + XLine 4)/4)


Satisfaction_train<- read.csv("satisfaction_train.csv", sep=",", header=T, stringsAsFactors = F)

Satisfaction_train<- Satisfaction_train[Satisfaction_train$Classification=="Seoul City",]

Satisfaction_train<- Satisfaction_train[,c(1,6)]

Satisfaction_train<- Satisfaction_train[c(order(Satisfaction_train$Period)),]

Satisfaction_train$Subway<- as.numeric(as.character(Satisfaction_train$Subway))

Transport_Average_scale<- transform(Transport_Average,

    Average = scale(Average)

)

Satisfaction_train_scale<- transform(Satisfaction_train,

Subway= scale(Subway)

)

ggplot(Transport_Average_scale, aes(x=Year, y=Average)) +

xlab('YEAR') +

ylab('RATE') +

geom_line(colour="blue") +

geom_line(aes(x=Satisfaction_train_scale$Period,y=Satisfaction_train_scale$Subway),colour="Green") +

geom_point(size=2, shape=19, colour="blue") +

geom_point(aes(x=Satisfaction_train_scale$Period,y=Satisfaction_train_scale$Subway),colour="Green") +

theme_bw() +

scale_x_continuous(breaks = c(2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016)) +

ggtitle("Utilization according to annual subway satisfaction") +

theme(plot.title = element_text(hjust = 0.5))
```

Using the ggplot function, we can see how subway utilization varies depending on subway satisfaction as shown in Figure 5. The average data, which means the average of the number of people transported by the subway in Seoul by year, and the subway variable, which means the subway satisfaction score of the statistical data on traffic use satisfaction in Seoul, were standardized and then visualized. Figure 5 shows that changes in subway satisfaction and subway utilization have a positive relationship.

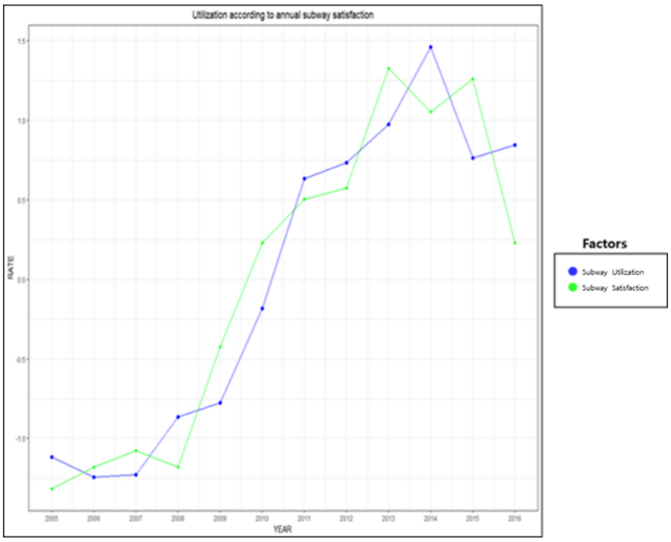

Fig. 5. ggplot Visualization

Since it was confirmed that the subway use rate has a correlation with satisfaction, related public data were used to analyze the factors affecting the subway use rate.

Table8. Statistical Data on Public Transportation Rate in Seoul

Period	Main(blue)
	Branch(Green)
		Grand(Red)	Circle(Yellow)	Village Bus	Subway	Medium-sized Taxi	Model Taxi

2005	800	1,400	600	500	800	1,900	4,500
2006	800	1,400	600	500	800	1,900	4,500
2007	900	1,700	700	700	900	1,900	4,500
2008	900	1,700	700	700	900	1,900	4,500
2009	900	1,700	700	700	900	2,400	4,500
2010	900	1,700	700	700	900	2,400	4,500
2011	900	1,700	700	700	900	2,400	4,500
2012	1,050	1,800	850	750	1,050	2,400	4,500
2013	1,050	1,800	850	750	1,050	3,000	5,000
2014	1,050	1,800	850	750	1,050	3,000	5,000
2015	1,200	2,300	1,100	900	1,250	3,000	5,000
2016	1,200	2,300	1,100	900	1,250	3,000	5,000

The period variables in Table 8 refer to the year 2005 to 2016, and other variables refer to each public transportation fee, and the Seoul Metropolitan Government's public transportation fee statistical data provided by the Seoul Open Data Square are refined as shown above.

Table 9. Data on the Crowdiness of Seoul Subway

Period	**Average**	**Line 1**	**Line 2**	**Line 3**	**Line 4**
2005	175	135	225	142	196
2006	171	128	221	137	189
2007	171	129	221	137	189
2008	156	122	202	122	172
2009	156	122	202	122	172
2010	167	144	196	149	180
2011	167	144	196	149	180
2012	166	144	202	147	169
2013	166	144	202	147	169
2014	152	106	192	134	176
2015	152	106	192	134	176
2016	152	106	192	134	176

The period variables in Table 9 are Year from 2005 to 2016, and the average variables are crowded, Line 1, Line 2, Line 3 and Line 4 of the Seoul Metro's congestion statistics provided by Seoul's Open Data Square.

Table 10. Air quality measurement data by subway station in Seoul

Year	**PM10**	**CO2**	**HCHO**	**CO**
2005	115.36	553.3	14.72	0.87
2006	113.12	566.38	13.83	0.91
2007	112.23	559.09	14.43	0.82
2008	95.17	511.67	15.46	0.82
2009	96.37	500.1	15.7	0.84
2010	92.16	507.79	17.02	0.71
2011	92.5	514.37	13.83	0.93
2012	90.2	522.3	18.8	1

2013	91	534	18.7	1
2014	90.36	569.45	18.91	1.05
2015	90.65	555.59	11.9	0.87
2016	84.1	561.5	12.9	0.9

The period variables in Table 10 refer to the 'year' from 2005 to 2016, the PM10 variable refers to fine dust concentration (μg/m3), the CO2 variable refers to carbon dioxide concentration (ppm), the HCHO variable refers to formaldehyde concentration (μg/m3), and the CO variable refers to carbon monoxide concentration (ppm) also The air quality measurement data by subway station in Seoul provided by Data Square in Seoul was refined as shown in the table above.

Table 11. Statistics on the comfort facilities in Seoul

Period	Elevator installation station	Number of Elevator Facilities	Escalator installation station	Number of escalator Facilities
2005	109	274	53	243
2006	109	275	57	259
2007	109	275	60	266
2008	109	275	62	276
2009	109	282	62	294
2010	113	298	77	384
2011	114	300	82	406
2012	115	309	84	424
2013	116	317	90	445
2014	116	329	93	490
2015	117	336	96	513
2016	116	330	98	524

The period variable in Table 11 is the year from 2005 to 2016, the number of stations variable is the number of stations with elevators and escalators in subway stations in Seoul, and the number of facilities variable is the total number of facilities with elevators and escalators installed in subway stations in Seoul. The statistical data of Seoul subway convenience facilities provided by the Seoul Open Data Center were refined as shown in the table above.

Table 12. Execution result source code

Execution result source code

```
Subway_Factor<- Transport_Average[,c(1,6)]

Fare <- read.csv("Fare.csv", sep=",", header=T)

Fare <- Fare[,c(1,7)]

Fare <- Fare[-c(1:4),]

names(Fare)[names(Fare) == "Public transportation fee.4"] <- c("cost")

Subway_Factor<- cbind(Subway_Factor, Fare$cost)

congested <- read.csv("congested.csv", sep=",", header=T)

congested <- congested[-c(1,2),c(1,4)]

congested <- congested[c(order(congested$Period)),]

congested <- congested[-c(1:5),]

Subway_Factor<- cbind(Subway_Factor, congested$Subway Congestion.1)

Airquality<- read.csv("Airquality_year.csv", sep=",", header=T)

Airquality<- Airquality[,-c(2)]

Airquality<- Airquality[-c(1,2,3),]

Airquality<- Airquality[c(order(Airquality$Year)),]

names(Airquality)[names(Airquality) == "Retention Cirteria"] <- c("PM10")

names(Airquality)[names(Airquality) == " Retention Cirteria.1"] <- c("CO2")

names(Airquality)[names(Airquality) == " Retention Cirteria.2"] <- c("HCHO")

names(Airquality)[names(Airquality) == " Retention Cirteria.3"] <- c("CO")

Subway_Factor<- cbind(Subway_Factor, Airquality$PM10, Airquality$CO2, Airquality$HCHO,Airquality$CO)

Elevator <- read.csv("Elevator.csv", sep=",", header=T)

Elevator <- Elevator[Elevator$division=="Seoul Metro" & Elevator$ division.1=="Subtotal",]

Elevator <- Elevator[,c(1,6,8)]

Elevator <- Elevator[c(order(Elevator$Period)),]

Elevator <- Elevator[-c(1),]

Subway_Factor<- cbind(Subway_Factor, Elevator$Elevator.1,
Elevator$Escalators.1)

    colnames(Subway_Factor) = c("Year", "Subway", "Fare", "Congested",
"PM10", "CO2", "HCHO", "CO", "Elevator", "Escalator")

    Subway_Factor[,3] <- as.integer(gsub(",", "", Subway_Factor[,3]))

    Subway_Factor$Congested<-
as.numeric(as.character(Subway_Factor$Congested))

    Subway_Factor$PM10 <-
as.numeric(as.character(Subway_Factor$PM10))

    Subway_Factor$CO2 <- as.numeric(as.character(Subway_Factor$CO2))

    Subway_Factor$HCHO<-
as.numeric(as.character(Subway_Factor$HCHO))

    Subway_Factor$CO<- as.numeric(as.character(Subway_Factor$CO))

    Subway_Factor$Elevator<-
as.numeric(as.character(Subway_Factor$Elevator))

    Subway_Factor$Escalator<-
as.numeric(as.character(Subway_Factor$Escalator))

    Subway_Factor_scale<- transform(Subway_Factor,

        Subway = scale(Subway),

        Fare = scale(Fare),

        Congested = scale(Congested),

        PM10 = scale(PM10),

        CO2 = scale(CO2),

        HCHO = scale(HCHO),

        CO = scale(CO),

        Elevator = scale(Elevator),

        Escalator = scale(Escalator)

    )

    Subway_Factor_scale<- Subway_Factor_scale[,-c(1)]

    Subway_Factor_scale

    install.packages("psych")

    library(psych)

    pairs.panels(Subway_Factor_scale[c("Subway", "Fare", "Congested",
"PM10", "CO2", "HCHO", "CO", "Elevator", "Escalator")])

    factor <- lm(Subway ~ ., data=Subway_Factor_scale)

    summary(factor)
```

Bus variables in Seoul's public transportation fare statistics data, average variables in Seoul's subway congestion statistics data, PM10, CO2, HCHO, and CO variables in air quality measurement data for each subway station in Seoul, and the number of elevator and escalator facilities in Seoul's subway convenience facilities statistical data. After standard normalization, using the pairs.panels function, we can check the factors that have a correlation with the subway usage rate as shown in Figure.

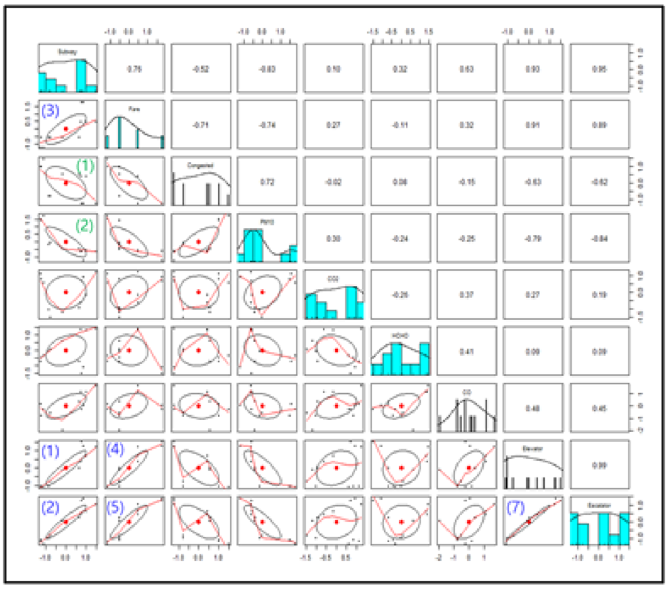

Fig. 6. pairs.panels Visualization

First, looking at Figure 6, from the top left, the dependent variables Subway: subway use rate, the remaining independent variables Fare: subway fare, Congested: subway congestion, PM10: fine dust concentration ($\mu g/m^3$), CO2: carbon dioxide (ppm), HCHO: Formaldehyde concentration ($\mu g/m^3$), CO: carbon monoxide concentration (ppm), Elevator: means the number of elevators, Escalator: means the number of escalators.

On the left-hand side of the diagonal reference of the graph, the positive relationship between the dependent and independent variables is that there is a correlation between the elements, whereas the negative relationship is that there is an inverse relationship between the elements. The semicircular has a weak correlation and the oval has a strong correlation.

On the right-hand side of the diagonal reference of the graph, the closer the value between the dependent and independent variables is 1, the more positive the correlation is, and the closer the zero the correlation is, the more irrelevant it can be interpreted. Conversely, the closer to –1, the more negative the relationship and the stronger the correlation is.

In Figure 6, elements that correlate strongly with subway utilization are the number of elevators and escalators, which are elements (1) and (2) of the blue-letter color. It can be interpreted that the number of elevators and escalators, i.e., subway usage, increases as convenience facilities on the subway increase. In addition, it has a weak positive relationship with the fare, which is the blue color element (3), and it can be interpreted that the increase in the usage rate even as the subway fare rises little by little is because the fare is reasonable compared to the average income of users.. In addition, elements (4) and (5) of the blue-letter color become the dependent variable and the independent variable becomes the number of elevators and escalators, which can be interpreted as the subway fare increases as amenities increase. If you look at the blue color (7), you can see the strongest positive correlation, and generally the elevators were also installed in the year when escalators were installed.

On the contrary, elements that have a negative relationship with subway utilization are the green color (1), the (2) element of subway congestion, and the concentration of fine dust. This can be interpreted as the subway congestion and fine dust increase, the subway utilization rate decreases.

Visualization through the pairs.panel function showed that congestion, fine dust concentration, and amenities were correlated with subway utilization. The next step is to generate a regression model using the lm function. Applying the summary function to a model created using the lm function yields the results shown in Table 12.

Table 13. Multiple Linear regression analysis result

-	**summary(factor)**				
Call	lm(formula = Subway ~ ., data = Subway_Factor_scale)				
Coefficients	variable	Estimate	stdError	T value	Pr(|t|)
	Fare	-2.014e-01	1.908e-01	-1.056	0.3687
	Congested	-6.547e-02	2.117e-01	-0.309	0.7773
	PM10	2.028e-01	6.465e-01	0.314	0.7743
	CO2	-1.787e-01	2.167e-01	-0.825	0.4701
	HCHO	9.125e-02	1.201e-01	0.760	0.5026
	CO	2.522e-01	8.047e-02	3.135	030519
	Elevateor	-1.973e-01	1.023e+00	-0.193	0.8594
	Escalator	1.363e+00	1.323e+00	1.030	0.3787
Residual standard error	0.1841 on 3 degrees of freedom				
Multiple R-squared	0.9908				
Adjusted R-squared	0.9661				
F-statistic	40.21 on 8 and 3 DF				
p-value	0.005733				

We perform an F-test to determine whether the regression model is statistically meaningful[3].

In a regression model, the R-squared value means the determinant of the model, and the larger the value, the higher the explanatory power. That is, regression models with near-zero determinant values are less useful, while regression models are more useful as the value of the determinant approaches 1. The value of R-squared below Table 12 is 0.9661, which can be determined to have a 96% explanatory power.

In the regression model, if the p-value value of F-Static is less than 0.05, the regression model is significant. Given the p-value values of F-Statistic under Table 12, which is the result of Figure 6, it can be determined that this regression model can be used to describe a significantly dependent variable, Subway utilization, as 0.005733 is less than 0.05.

V. ANALYSIS RESULTS

1. Extension of amenities

First of all, the analysis results of the data covered in this study show that subway usage increases/decreases proportionally as subway satisfaction increases/decreases in Figure 5.

When the final results confirm the factors that increase the satisfaction level, it was selected that the satisfaction level of the factors most closely related to utilization should be increased. As a result, we could see that the utilization rate increased simultaneously as the number of elevators and escalators corresponding to convenience facilities increased, which led to the need to expand the space of convenience facilities or increase the number of facilities.

2. Improvement of air quality in stations by establishing ventilation facilities

Next, based on the subway utilization rate, which is a dependent variable, the deterioration of air quality was inversely proportional to the increase in subway utilization when using the results of multiple linear regression analysis for each element. First of all, it was deemed necessary to improve fine dust and carbon dioxide, which accounted for the largest figures in the air quality data in the region, and air quality accounted for a large portion of the subway utilization rate.

As a priority, it also concluded that additional ventilation facilities should be deployed from the history of high concentrations of substances such as fine dust and carbon dioxide, which can negatively affect users of air quality in the region.

VI. CONCLUSION

In this paper, the data provided by the Seoul Open Data Square Portal was inspired by the government's recent free public transportation policy for Seoul City, which will play a major role in reducing fine dust by 0.8 tons. Factors that can increase public transportation satisfaction rate and utilization rate are analyzed using multi-linear regression analysis, and the correlation between utilization rate and each factor is confirmed. However, the satisfaction and utilization of buses and subways, which account for the largest portion of public transportation, were increasing, while the utilization of subways was falling sharply. Therefore, if the use of the subway without fossil fuels is increased, it will have a more positive effect on the reduction of fine dust, and analyzed factors that could be related to the subway utilization rate. As a result, the most relational factors analyzed were the number of convenience facilities and air quality in the region.

As the number of elevator and escalator facilities corresponding to subway amenities increased, the utilization rate of subways also increased at the same time. On the other hand, among the negative factors, carbon dioxide or fine dust, which corresponds to air quality in the region, greatly reduced the utilization rate, so history with particularly high levels of fine dust and carbon dioxide could conclude that the improvement of ventilation facilities is urgent.

It is predicted that the more data through other surveys and surveys, the more relevant factors will be found, and the effect of reducing fine dust will increase accordingly.

REFERENCES

[1] Kwang-Joo Moon, Hyeok-Gi Cheo, Kwon-Ho Jeon, Xiaoyang Yang, Fan Meng, Dai-gon Kim, Hyun-Ju Park, Jeong-Soo Kim, “Review on the Current Status and Policy on PM2.5 in China,” Journal of Korean Society for Atmospheric Environment, Vol. 34, No. 3, pp. 373-392, 2018.6.

[2] Kyung-Su Jang, Jun-Ho Yeo, “The Effects of Korea and Chinese Economic Growth on Particulate Matter in Korea: Time Series Cointegration Analysis,” Journal of Environmental Policy and Administration, Vol. 23, No. 1, pp. 97-117, 2015.3.

[3] Kuk-Hyung Lee, “A Study on Perception for Public Safety of Seoul Citizens using Multiple Regression Analysis,” The Institute of Internet, Broadcasting and Communication, Vol. 18, No. 1, pp. 195-201, 2018.

[4] Uh-Soo Kyun, Sung-Hoon Cho, Jeong-Joon Kim, Young-Gon Kim, “A Study on Perception for Public Safety of Seoul Citizens using Multiple Regression Analysis,” Journal of The Institute of Internet, Broadcasting and Communication, Vol. 18, No. 1, pp. 195-201, 2018.2. DOI: https://doi.org/10.7236/JIIBC.2018.18.1.195

[5] Seong-Eun Yang, Chang-Yeol Choi, Hwang-Kyu Choi, “Design and Implementation of Vehicle Route Tracking System using Hadoop-Based Bigdata Image Processing,” Journal of Digital Contents Society, Vol. 14, No. 4, pp. 447-454, 2013.12.
DOI: http://dx.doi.org/10.9728/dcs.2013.14.4.447

[6] Eun-Ju Park, Hye-Jin Choi, So-Jeong Park, So-Hyun Oh, Yong Lee, Jun-Ho Shim, “Efficient Processing of Multiple Group-by Queries in MapReduce for Big Data Analysis,” Journal of KIISE Transactions on Computing Practices, Vol. 21, No. 5, pp. 387-392, 2015.5.

[7] Jong－Ki Lee, “A Case Study on Practical Accounting Processing of Big Data Using R Programming,” Journal of Korean Computers and Accounting Review, Vol. 13, No. 1, pp. 1-22, 2015.6.

[8] Sang-Ho Choo, “Analyzing Factors Affecting Satisfaction of Public Transit Users,” Seoul City Research, Vol. 13, No. 3, pp. 65-78, 2012.9.

[9] Dae-Sung Son, “A Study on how to respond to fine dust using the foundation of Big Data,” Journal of Korea Intelligent Information Systems Society, pp. 23-24, 2017.11.