Nadia AL-Rousan    ORCID iD: 0000-0001-8451-898X

# Data Analysis of Coronavirus CoVID-19 Epidemic in South Korea Based on Recovered and Death Cases

Nadia AL-Rousan, Hazem AL-Najjar

Department of Computer Engineering, Faculty of Engineering and Architecture

Istanbul Gelisim university, Istanbul, Turkey

## Abstract

Coronavirus epidemic caused announcing emergency case in South Korea. The virus started with one infected case by January 20, 2020, where 9583 announced cases were reported by March 29, 2020. This indicates that the number of confirmed cases is increasing rapidly, which can cause national crises for South Korea. The aim of this study is to fill a gap between previous studies and the current development of CoVID-19 spreading, by extracting a relationship between independent variables and dependent variable. This research statistically analyzed the effect of sex, region, infection reasons, birth year, and released or diseased date on the reported numbers of recovered and deceased cases. The results found that sex, region, and infection reasons affected on both recovered and deceased cases, while birth year only affected on deceased cases. Besides, no deceased cases are reported for released cases, while 11.3% of deceased cases positive confirmed after their deceased. Unknown reason of infection is the main variable that detected in South Korea with more than 33% of total infected cases.

**Keyword-** Epidemiology, engineering and technology, infection, South Korea

## 1. Introduction

The first case of coronavirus CoVID-19 disease in South Korea is announced on 20th of January 2020[1]. The distance from South Korea (located on 37° North and 127° East) to China is 2123 kilometers. South Korea is considered as the third infected country by the epidemic of coronavirus after China and Italy[2]. Coronavirus infected more than 10156 cases by April 5, 2020 in South Korea while the global number of infected cases is 1201943 cases[3]. Referring to World Health Organization (WHO), there are 249127 recovered cases and 64781 deceased cases globally. South Korea has 183 deceased cases and 6463 recovered cases[4-5]. By time, the number of confirmed infected cases has been rapidly increased in South Korea. The growth rate of confirmed cases was rapid until March 11, 2020, while there was a slow increase after March 11, 2020 until the current time.

On the other hand, the first deceased case was reported on 20th of February to reach 183 cases by the end of April 5, 2020, while the first recovered case was reported on 7th of February to reach 6463 cases by the end of April 5, 2020. All of the infected cases suffered from several symptoms before confirming their infection by coronavirus disease[6-7], these symptoms started by feeling cold, flu, and pneumonia. Coronavirus test was made to around 10206 cases by the end of 8th of March in South Korea. It was reported that most of the infected cases in South Korea have visited some local cites before positive confirming their disease (i.e., isolation hospital, airport, restaurant, market, café, clinic, company, Movie Theater, etc.) [8].

Evidently, several probable reasons for spreading coronavirus in South Korea are summarized by many researchers in the field[9-10]. Several researches were conducted to find the probable reasons of spreading the coronavirus in South Korea rather than other countries. Researchers have started to extract information about the infected cases and analyzed the biomedical information and their medical histories to extract the main parameters that could cause coronavirus spreading. Researchers suggested that the spreading of coronavirus could be connected with sex, birth year, or the region they come from.

The aim of this research is to study the effect of several attributes on the spreading of coronavirus CoVID-19 in South Korea based on real collected data and published reports. The main target is to study the effect of sex, birth year, the region they come from, and the place they visited on the number of deceased and recovered cases in South Korea. Chi-Square Test is used to find the impact of the previous attributes on the number of recovered and deceased cases. The study would give an overview about the current situation in South Korea, besides, it may show the main parameters that can be used to build a forecasting models.

## 2. Methodology

As explained earlier, this research studies the effect of sex, age, region, and transportation on susceptibility to CoVID-19 in South Korea. Official time series data from Korea Centers for Disease Control and Prevention (KCDC) for Coronavirus disease 2019 (COVID-19) cases in South Korea from 20th of January until 29th of March are used[11]. The obtained data contains several information about 2771 infected cases (where the rest of data is missing and not reported) in South Korea namely, sex, birth year, the original country they come from, the region that they live in, whether they carrying any kind of disease before, infection reason and order, confirmed date, deceased or released date, and they current state. The data contain several missed variables that excluded from the analysis to give a clear overview about coronavirus epidemic in South Korea.

Both statistical analysis and Chi-Square Test are used to analyze the collected data and to ensure about the impact of sex, region, infection reason, released and deceased cases, and the birth year on the number of recovered and deceased cases in South Korea. Chi-Square test is heavily recommended to be used in survey research, business intelligence, engineering and scientific research as well. A chi-square test ($\mathrm{X}^2$) is a common mathematical test that used to check the relationship between two variables in a contingency table that present (multivariate) frequency distribution of the variables. Chi-Square test which is normally known as independence test is used to test the hypotheses for categorical variables and to test whether these variables are independent population variable. Chi-Square test can be calculated by finding the summation of dividing the square difference between the observed (O) and the expected (E) values by the expected value for each category in data as shown in Formula 1[12].

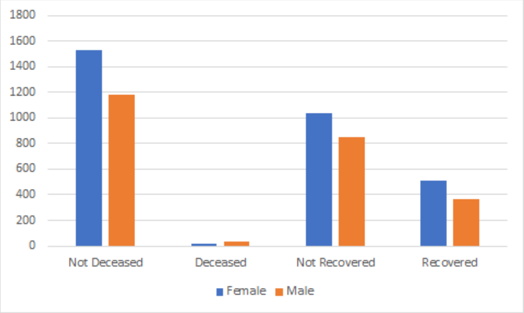
$$\sum X_{i-j}^2 = \frac{(O-E)^2}{E} \quad (1)$$

where, X2 denotes the Chi-Square test value, O and E denote the observed and the expected values respectively. A significant test should be done in order to determine whether the data is significant or not. The variables are significant if the probability to chance of association occurrence (P) not more than one out of 1000 cases (0.0001), otherwise, it will be considered as not significant. Thus, the association between variables will be rejected. Moreover, to understand the relationship between independent variables and recovered and deceased cases, a cross tabulation method is used. The cross-tabulation method is used between one independent and one dependent variable, to understand how the dependent variable is moving based on the movement of independent variables. Besides, multinomial logistic regression classification method is used to check the validity and the robustness of using Chi-Square test to trace the association of each variables namely, sex, confirmation date, birth year, region, and infected reason on both recovered and deceased cases.

## 3. Results

Studying the effect of sex variable on the recovered cases found that 2765 cases are classified to 1547 female cases and 1218 male cases. The number of recovered female cases is 511 cases while 366 recovered male cases exist as shown in Figure 1.

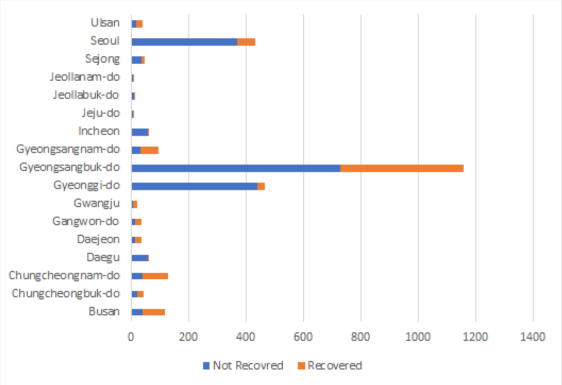
**Figure 1:** Distribution of recovered, unrecovered, deceased, and un-deceased cases based on sex.



By using Chi-Square Test to find the impact of sex variable on the number of recovered cases, the results found that Chi-Square function is $X^2$ (2, 2771) = 14.44, p=0.006 which indicates that sex variable is statistically significant with the number of recovered cases. The same test was used to find the effect of sex variable on the number of deceased cases. The results found that Chi-Square function is $X^2$ (2, 2771) = 12.64, =0.002. The results indicate that sex variable is significant with the number of deceased cases as well. The number of deceased male cases is greater than the number of deceased female cases with 36 and 17 cases respectively as shown in Figure 1.

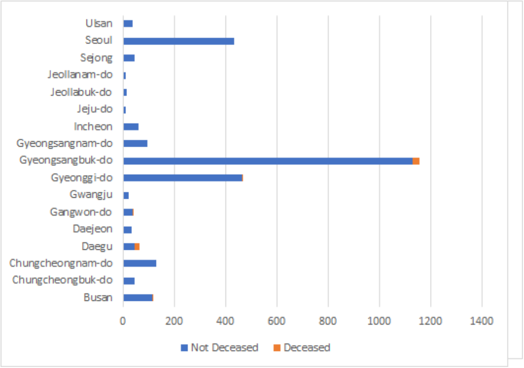
Moreover, studying the effect of region on the number of recovered and deceased cases found that 879 cases were classified as recovered and the rest of patients either deceased or isolated. The analysis found that Gyeongsangbuk-do registered highest number of recovered cases(i.e., 430 cases). Besides, the maximum number of infected cases were reported in Gyeongsangbuk-do region, while the range ratio of recovered cases to infected cases is from 3% to 70%. as shown in Figure 2.

**Figure 2:** Ratios of recovered cases to infected cases based on region.



The results of Chi-Square Test are defined as $X^2$ (16, 2771) = 516.49, p<.0001 which indicates that region variable is statistically significant with the number of recovered cases, where the Chi-Square result of deceased cases is defined as $X^2$ (16, 2771) = 326.20, p<.0001. In addition, it is found that none of unspecified cases is reported as deceased, while 2.6%, 31.7%, 2.7%, 0.21% and 2.4% of cases in Busan, Daegu, Gangwon-do, Gyeonggi-do, Gyeongsangbuk-do regions are reported as deceased cases, respectively. Figure 3 shows the ratios of deceased cases to infected cases in all regions.

**Figure 3:** Ratios of deceased cases to infected cases based on region.



Moreover, to understand the impact of the infection reasons on both studied variables including deceased and recovered cases, the study analyzed the infection sources types using cross-tabulation and Chi-Square test. The collected dataset from south Korean hospital classified the infection reasons into several groups namely, Bonghwa Pureun Nursing Home (F1), Changnyeong Coin Karaoke (F2), Cheongdo Daenam Hospital (F3), contact with patient (F4), Dongan Church (F5), ETC (F6), Eunpyeong St. Mary's Hospital (F7), Geochang Church(F8), Guro-gu Call Center(F9), Gyeongsan Cham Joeun Community Center(F10), Gyeongsan Jeil Silver Town (F11), Gyeongsan Seorin Nursing Home (F12), gym facility in Cheonan (F13), gym facility in Sejong (F14), Ministry of Oceans and Fisheries (F15), Onchun Church (F16), overseas inflow (F17), Pilgrimage to Israel (F18), River of Grace Community Church (F19),Seongdong-gu APT (F20), Shincheonji Church (F21), Suyeong-gu Kindergarten (F22) and Unkown (F23). Studying the infection reasons in both recovered and deceased cases found that 928 cases are recorded without determining the reason of infection. The analysis found that 409 recovered cases and 38 deceased cases are recorded in south Korean hospitals as unknown reason of infection. Besides, it is found that direct contact contact with patient, ETC, overseas inflow, Guro-gu Call Center are the main four causatives to be infected by the virus. The results found that all the cases that infected because of the following reasons are recovered, the reasons are Changnyeong Coin Karaoke, Pilgrimage to Israel, River of Grace Community Church, and Suyeong-gu Kindergarten. The range of recovered percentages of patients is between 3.57% and 66.67% of recovered cases based on infection case, respectively, where Guro-gu Call Center, Bonghwa Pureun Nursing Home, Dongan Church, Gyeongsan Jeil Silver Town. Gyeongsan Seorin Nursing Home, and Gyeongsan Cham Joeun Community Center showed no recovering in the infected patients as shown in Figure 4. In addition, 72% and 28% of deceased cases are infected because of unknowing reason or one of suggested cases from F1 to F22. Moreover, no deceased case until March 29, 2020 is reported because of visiting Wuhan. The ratio of deceased cases is drawn in Figure 5.

**Figure 4:** Ratio of recovered cases to infected cases based on the infection reasons

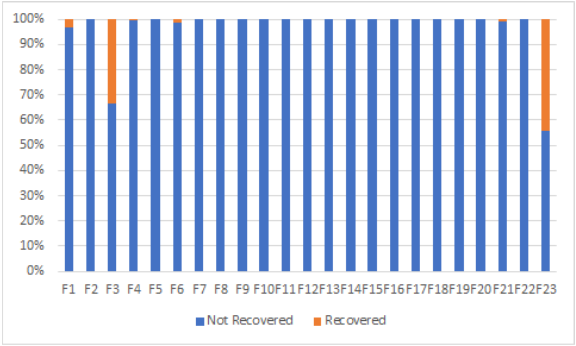
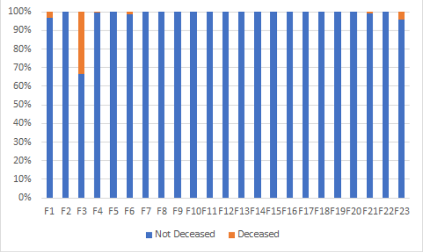


**Figure 5:** Ratio of deceased cases to infected cases based on the infection reasons



The results of Chi-Square of infection reasons on the number of recovered and deceased cases are $X^2$ (22,2771) = 383.5, p<.0001, and $X^2$ (22,2771) = 158.23, p<.0001, respectively, which indicates that the infection reasons are statistically significant with the number of recovered and deceased cases.

In addition, to validate the speed of south Korean hospitals in testing the Coronavirus cases, a relationship between released and deceased cases are studied. The results found that no deceased cases were reported for any of the recovered cases, where 11.3% of 53 deceased cases were confirmed as coronavirus infected cases either after their deceased or in the same day of their deceased, which indicates that the process to discover coronavirus symptoms is acceptable. Moreover, by studying the relationship between the birth year and the number of recovered and deceased cases, the results found that no relationship was found between the birth year and the number of recovered cases, while 83% of deceased cases are for persons have more than 60 years old. Birth date is effective variable on the number of deceased and recovered cases the Chi-Square test found that the relationship is statistically significant with Chi square function equal to $X^2$ (99, 2771) = 327.89, p<.0001 and $X^2$ (99, 2771) = 175.32, p<.0001, respectively.

To verify the relationship between selected independent variables and one of the dependent variables( i.e., recovered and deceased), a multinomial logistic regression was used. The overall percentage results of percent correct for deceased and recovered cases are 99.5% and 88.0%, respectively as shown in Table 1. In addition, to check the significance level of the independent variables, a likelihood test is adopted for both classifiers. The deceased results showed that Birth_Date, Sex, Country, Region, Infection_Reason and confirmed_date are statistically significant factors to predict the deceased cases. The recovered results showed that Sex, Region, Infection_Reason, confirmed_date and Birth_Dateare statistically significant factors to predict the recovered cases, where Country is not statistically significant. The results revealed that determining the infection reason and confirmed date are useful information to determine the deceased cases, besides the results found that determining the region of the patients, early detecting the COVID-19, the reason of infection, and the gender of the patient could increase the probability of treating the patients.

**Table 1:** classification results based on multinomial logistic regression

| | | **.00** | **1.00** | **Percent Correct** |
|---|---|---|---|---|
| **Death** | .00 | 2361 | 3 | 99.9% |
| | 1.00 | 8 | 42 | 84.0% |
| | Overall Percentage | 98.1% | 1.9% | 99.5% |
| **Recovered** | .00 | 1426 | 212 | 87.1% |
| | 1.00 | 78 | 698 | 89.9% |
| | Overall Percentage | 62.3% | 37.7% | 88.0% |

**Table 2:** Likelihood Ratio Tests

| | | Model Fitting Criteria | | | Likelihood Ratio Tests | | |
|---|---|---|---|---|---|---|---|
| | **Effect** | **AIC of Reduced Model** | **BIC of Reduced Model** | **-2 Log Likelihood of Reduced Model** | **Chi-Square** | **df** | **Sig.** |
| **Death** | Intercept | 482 | 1703 | 60 | 0 | 0 | |
| | Birth_Date | 574 | 1790 | 154 | 94 | 1 | .000 |
| | Sex | 487 | 1703 | 67 | 7 | 1 | .006 |
| | Country | 17711 | 18243 | 17527 | 17467 | 119 | 0.000 |
| | Region | 846 | 2027 | 438 | 379 | 7 | .000 |
| | Infection_Reason | 493 | 1593 | 113 | 54 | 21 | .000 |
| | confirmed_date | 449 | 1352 | 137 | 77 | 55 | .027 |
| **Recovered** | Intercept | 1606 | 2827 | 1184 | 0 | 0 | |
| | Sex | 1632 | 2848 | 1212 | 28 | 1 | .000 |
| | Country | 1604 | 2820 | 1184 | 1 | 1 | .443 |
| | Region | 1738 | 2270 | 1554 | 370 | 119 | .000 |
| | Infection_Reason | 1639 | 2820 | 1231 | 47 | 7 | .000 |
| | confirmed_date | 1625 | 2725 | 1245 | 61 | 21 | .000 |
| | Birth_Date | 1917 | 2820 | 1605 | 422 | 55 | .000 |

## 4. Discussion

Sharing the patients' information can help researchers and governments, to understand the virus transmission. The sequence of CoVID-19 and the virus are shared between different laboratories to study the virus and to find characterizes of the virus. In this article, we analyzed the patient's information after removing the missing of each variable separately. This study adopted the following variables including sex, region, infection reasons, birth date, confirmed- deceased date and confirmed- recovered date. The study found that sex has a strong relationship with recovered and deceased cases, besides the majority of infected patients are male. This conclusion is in line with the the conclusion in[13], which proved that the number of female smokers is less than the number of male smokers. The region variable is used with recovered and deceased cases, the results revealed that the number of deceased and recovered are changed based on the region of the infected case. This result is in line with findings [9] that showed the number of cases is changed based on the weather variables, geographical area, populous density, people communication, transports, and the nature of Korians' bodies that permitted the incubation of the disease. The infection reasons classification can give a hint to researchers in the field, to trace the reason of the infection to classify the COVID-19 based on the way of the infection, which may help the doctors to give extra treatment to special cases. The infection reasons variable gave an alert to all researchers and governments that the virus has a strong transmission ability between people, besides there is a limit information about how the virus can infect new case without communicating with infected case. The birth date variable showed a good prove that the majority of deceased cases is more than 60 years old, where no indicator found between the number of recovered cases and birth date. The results are in line with finding of the China Center for Disease Control, which indicate that the fatality rate of infected cases less than 40 years is less than patients over 80 years.

Moreover, the results of confirmed- deceased date showed that the CoVID-19 test of 11.3% deceased cases are reported either after deceased or in the same day of the deceased. This revealed that the doctors were unable to treat these cases since no positive or negative information about the virus infection is reported, where all the recovered case did not show any deceased cases until March 29, 2020. Besides, the results found that multinomial logistic regression could give initial indicator about the possibility to survive or to deceased based on the collected data. It is found that the results of multinomial logistic are in line with the results of Chi-Square test.

## 5. Conclusion

This research highlighted the main variables that could be considered to understand the CoVID-19. The main variables that considered in this research are sex, region, infection reasons, birth date, confirmed-deceased date and confirmed- recovered date. After discussing the obtained results, it is found that to mitigate the coronavirus disease in South Korea several procedures should be followed. These procedures related to prevent any direct contact with patients especially those inside isolated hospitals and prevent any kind of community events (i.e. Visiting patients, go to restaurants, shopping at groups and big stores, etc.). Besides, the processes of testing people against coronavirus should be faster. Keeping South Korea clean from coronavirus would affect on all nearby countries.

## Conflict of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Funding Information

The authors received no specific funding for this work.

## References

1. Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. International journal of antimicrobial agents, 105924.
2. 2019-nCoV Global Cases (by Johns Hopkins CSSE). Available online: https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 (accessed on 11 Mrch 2020). Google Scholar
3. World Health Organization. (2020). Home care for patients with suspected novel coronavirus ( nCoV) infection presenting with mild symptoms and management of contacts: interim guidance.
4. Shim, E., Tariq, A., Choi, W., Lee, Y., & Chowell, G. (2020). Transmission potential of COVID-19 in South Korea. medRxiv.

5. Giovanetti, M., Benvenuto, D., Angeletti, S., & Ciccozzi, M. (2020). The first two cases of 2019-nCoV in Italy: where they come from?. *Journal of Medical Virology*.
6. European Centre for Disease Prevention and Control (ECDC). Risk assessment: Outbreak of acute respiratory syndrome associated with a novel coronavirus, Wuhan, China; first update 2020 [updated 22 January 2020]. Available from: https://www.ecdc.europa.eu/en/publications-data/risk-assessment-outbrea
7. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y.,... & Yu, T. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet.
8. Khan, N., & Naushad, M. (2020). Effects of Corona Virus on the World Community. Available at SSRN 3532001.
9. Al-Rousan, N., & Al-Najjar, H. (2020). Nowcasting and Forecasting the Spreading of Novel Coronavirus 2019-nCoV and its Association With Weather Variables in 30 Chinese Provinces: A Case Study. Available at SSRN 3537084.
10. Yoo, J. H., & Hong, S. T. (2020). The outbreak cases with the novel coronavirus suggest upgraded quarantine and isolation in Korea. Journal of Korean Medical Science, 35(5).
11. Korea Centers for Disease Control and Prevention (2020). http://ghdx.healthdata.org/organizations/korea-centers-disease-control-and-prevention-kcdc
12. Islam, J. Y., Khatun, F., Alam, A., Sultana, F., Bhuiyan, A., Alam, N.,... & Nahar, Q. (2018). Knowledge of cervical cancer and HPV vaccine in Bangladeshi women: a population based, cross-sectional study. BMC women's health, 18(1), 15.