Statistical investigation of COVID-19 spread: a regression analysis

Macole Sabat[1], Amany Fayssal Chahine[2,3], Mira Sabat[4,a]

[1] Mechanical Engineering Department, Faculty of Engineering, University of Balamand, Lebanon, macole.m.sabat@balamand.edu.lb

[2] School of Veterinary medicine, Faculty of health and medical sciences, University of Surrey, United Kingdom

[3] Department of Biology, Faculty of Arts and Sciences, University of Balamand, Lebanon, a.chahine@surrey.ac.uk

[4] Department of Mathematics, Faculty of Arts and Sciences, University of Balamand, Al Kurah, Lebanon, mira.sabat@balamand.edu.lb

[a] Corresponding author

Abstract

This paper presents a statistical investigation of the COVID-19 spread in 31 countries from 31 December 2019 to 26 March 2020. It shows that the total number of infected persons follows the same trend in almost all countries using a regression analysis and they fit a third-degree polynomial. The obtained profiles are validated with the number of infections in 26 and 27 March with an estimate of the percentage error. These theoretical profiles can be used to estimate the number of infections in the sharp increase phase. In addition, a grouping of the different countries was put in place. This is done based on an estimate of the Poisson parameter and on the sign of the leading coefficient of the fitted polynomial. The obtained results may help the countries in the early phases to compare to their similar counterparts and take the necessary measures to limit the virus spread.

1. Introduction

COVID-19 is an infectious disease caused by a novel virus belonging to the coronaviruses family. This family includes a number of viruses that infect humans, so called, common human coronaviruses and others attacking animals such as bats, snakes etc…(Gorbalenya et al. 2020). COVID-19 is an enveloped positive stranded RNA virus that infects vertebrates (Abdulamir & Hafidh, 2020). It has been discovered for the first time in December 2019, in Wuhan, China from which it has rapidly spread to almost all the countries in the world (Abdulamir & Hafidh, 2020). As of 25 March 2020, the worldwide confirmed cases were 416686 with 18589 deaths in 197 countries, territories and regions (WHO, March 25 2020). Up of 85% of the infected people have developed mild to moderate infection, 10% with severe infection and about 5% of critical illness half of which died (Abdulamir & Hafidh, 2020). The death was deeply correlated to the age (Deng et al., 9000) and the comorbidities such as hypertension and cancer (Tian et al.). Unfortunately, until now, the pathogenesis of this infection is still unclear.

Therefore, detecting similarities between the spread of the virus in different regions may help the scientists better understand how this virus acts and helps predicting its behavior. Based on this, the objectives of this study are:

- to find the parameters that can be used in order to categorize the countries that might end up having similar trends,
- to detect these trends through regression analysis.

2. Methodology

The number of infected persons (ECDPC, 26 March 2020) from 31 December 2019 to 25 March 2020 is scaled to get the number of infections per million population of each country and plotted versus the number of days of the virus spread. This is done for 31 countries. Successive least square regression analysis is conducted to determine a standard threshold above which the curves of different countries show similarities. This threshold is found to be the day on which the total number of confirmed cases reaches 0.2 per million population. The least square polynomials are then determined for each country and the results are validated and interpreted.

The resulting curves are fitted with third degree polynomials. Three forth of which shows R^2 values greater than 0.98 and 58% greater than 0.99.

The least square polynomials are validated by comparing the expected total cases on 27 and 28 March computed based on these polynomials to the actual total cases (ECDPC, 28 March 2020).

The countries are grouped using 2 criteria:

- the Poisson parameter, λ: average number of infections per unit day;
- the sign, a ,of the leading coefficient in the fitted polynomial.

Therefore, seven groups were found:

- Group 1: $\lambda > 20$: Italy, Switzerland, Spain, Austria
- Group 2: $5 < \lambda \leq 20$, $a > 0$: Germany, France, Netherland, USA, UK
- Group 3: $5 < \lambda \leq 20$, $a < 0$: Norway, Iran, Denmark, Bahrain, Qatar

- Group 4: $1 < \lambda \leq 5$, $a > 0$: Chile, Lebanon, Kuwait, Singapore, Australia, Canada
- Group 5: $1 < \lambda \leq 5$, $a < 0$: Greece, South Korea, Turkey
- Group 6: ≤ 1 , $a > 0$: UAE, Tunisia, Brazil, Morocco, Philippines
- Group 7: ≤ 1 , $a < 0$: Peru, Sri Lanka, China.

3. Results and Discussions

The actual data and the least square third degree polynomial is sketched for the different countries of a given group. The equation of the polynomial and the R^2 values are displayed on the graph as well. In addition, Table 1 represents the actual number of infections per million population, the expected number of infections per million population and the percentage error for the two days following the regression: March 27 and March 28, 2020.

3.1 Group 1

The values of R- square are all greater than 0.995 as shown in Figure 1. Thus in this group the 3rd order polynomial is a very good fit for the sharp increase in the total confirmed cases. The curves of Switzerland, Span and Austria follow a very similar trend. In fact, the coefficients a of these three curves have the same order of magnitude $M(a) = 0.1$. Italy, however, has a spread speed that is reduced by a factor of 0.4 compared to the other countries in group 1.

Concerning the estimates of Poisson parameter,λ, it is 22.17, 35.14, 36.18 and 40.71 respectively for Austria, Spain, Italy and Switzerland. Until 25 March, the most severe increase is in Switzerland as having the highest

Poisson parameter. The smallest errors in this category are the once obtained from the data of Austria as shown in Table 1.

Table 1: Percentage error for the different countries

		27 March	28 March
	Country	**%Error**	**%Error**
Group 1	Italy	5.33	7.52
	Switzerland	6.93	6.99
	Spain	6.01	6.01
	Austria	3.00	1.22
Group 2	Germany	2.84	3.56
	France	2.84	3.56
	Netherland	1.91	3.77
	USA	6.40	9.03
	UK	7.96	15.30
Group 3	Norway	2.52	7.12
	Iran	7.36	11.87
	Denmark	12.66	19.26
	Bahrain	6.66	3.27
	Qatar	18.65	29.62
Group 4	Chile	1.26	3.99
	Lebanon	0.57	3.87
	Kuwait	9.74	8.23
	Australia	22.96	20.45
	Canada	29.42	31.45
	Singapore	0.33	13.68
Group 5	Greece	0.63	0.83
	South Korea	6.55	9.96
	Turkey	21.76	41.58
Group 6	UAE	21.25	29.97
	Tunisia	18.08	11.64
	Brazil	1.64	0.19
	Morocco	0.04	1.26
	Philippines	7.10	9.18
Group 7	Peru	2.17	4.19
	Sir Lanka	0.78	0.11
	China	6.78	9.03

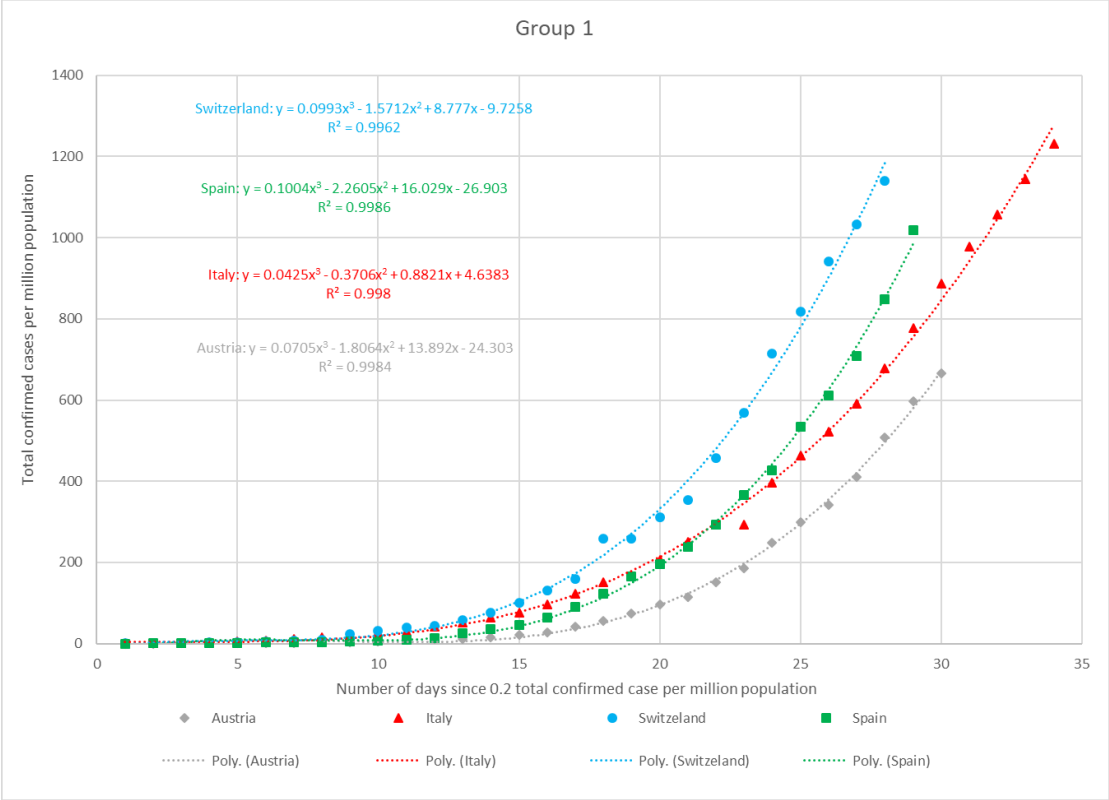


Figure 1: Group 1

3.2 Group 2

The values of R^2 are all greater than 0.992 as shown in Figure 2. Concerning the estimates of Poisson parameter,λ, it is 5.11, 7.81, 12.53, 14.31 and 14.67 respectively for UK, USA, France, Netherland and Germany. The smallest errors in this category are the once obtained from the data of Netherland as shown in Table 1. The order of magnitude of the coefficients a for the cases of Netherland, Germany and USA is 0.04, leading to similar trends of these curves. France and UK on the other hand, are similar with 0.02 as the order of magnitude of a, leading to an expected decrease in the speed of spread of the virus by half compared to the other 3 countries of this group.

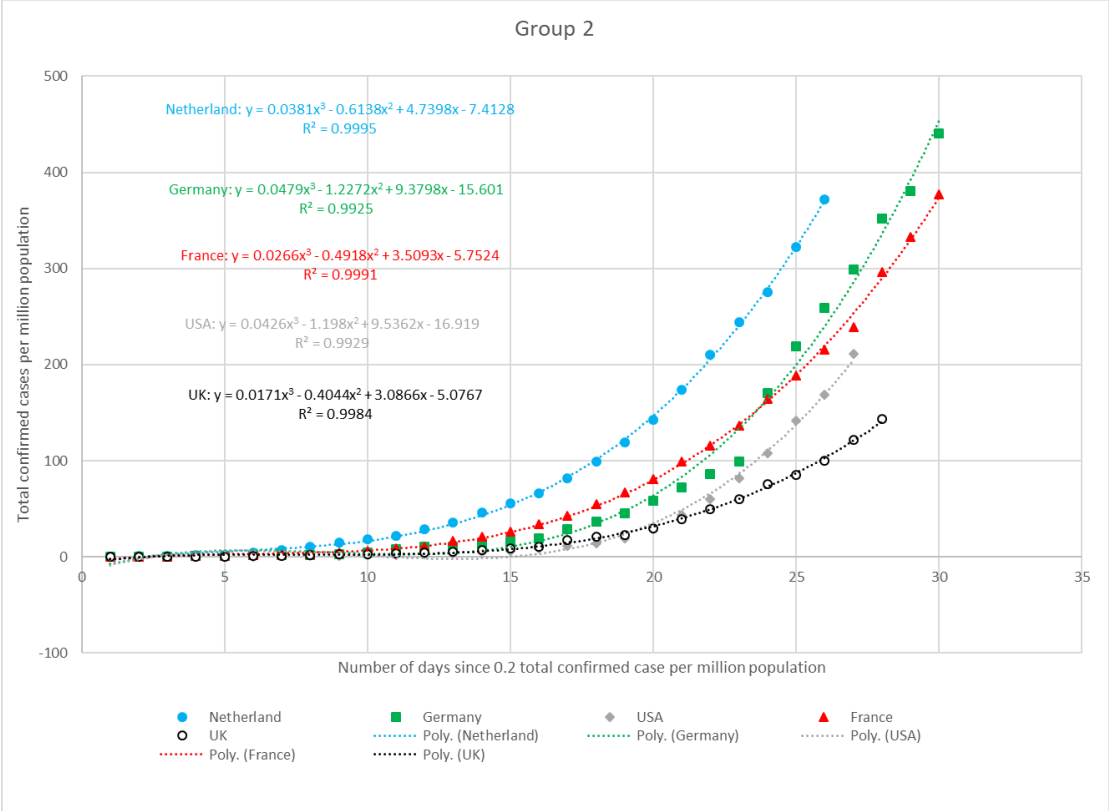


Figure 2: Group 2

3.3 Group 3

Even though, the R^2 values are all greater than 0.95 as shown in Figure 3, the data of Qatar and Denmark is not following a clear logical trend compared to the other countries and groups, thus the analysis of this data is not reliable which also explains the corresponding high percentage errors. As for Norway, Iran and Bahrein, the resulting curves can be better fitted with polynomial of degree two as follows:

- Norway: y = 0.9424x2 - 7.7811x + 14.222, R^2 = 0.9962;
- Iran: y = 0.2913x2 + 0.0332x - 7.7945, R^2 = 0.9966;
- Bahrein: y = 0.239x2 + 0.2623x + 6.7634, R^2 = 0.9843.

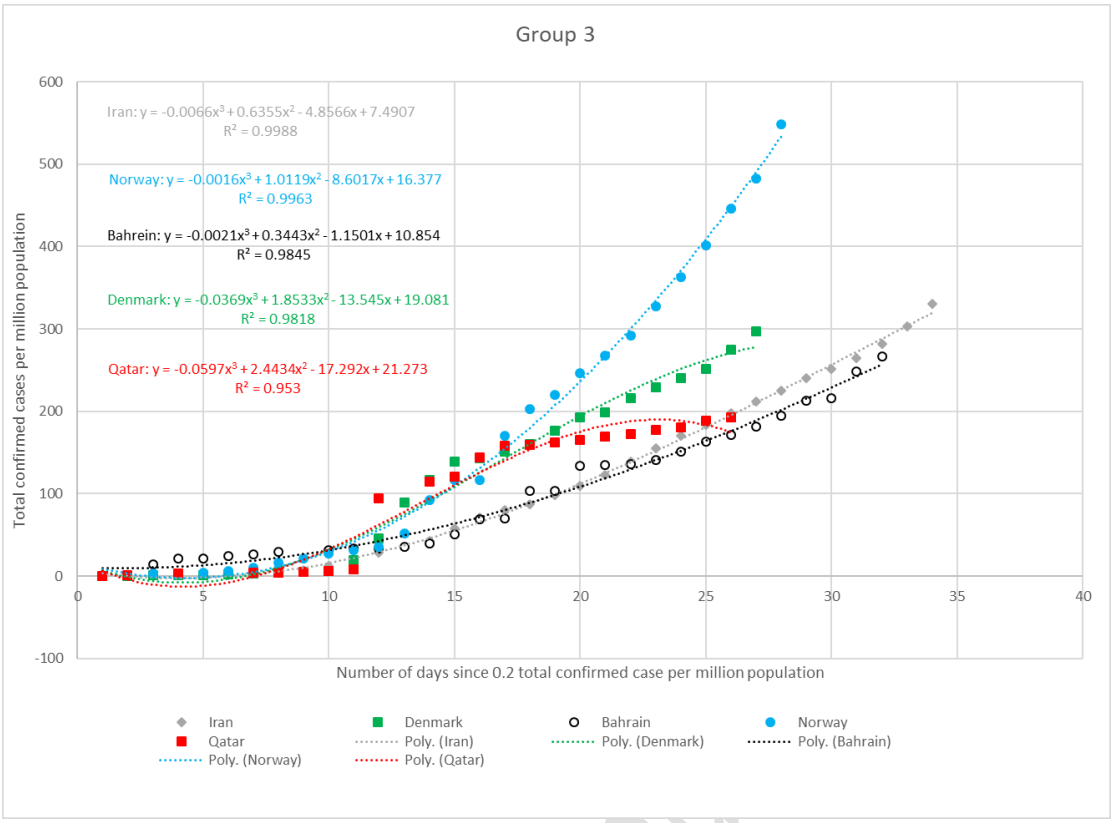


Figure 3: Group 3

3.4 Group 4

The percentage errors in this category are only good for Chile and Lebanon. It is important to note that although Australia and Singapore were among the first countries to reach the 0.2 per million population threshold, they are still in the phase of sharp increase unlike South Korea and China that have a more or less stabilized spread over such long period (see groups 5 and 7).

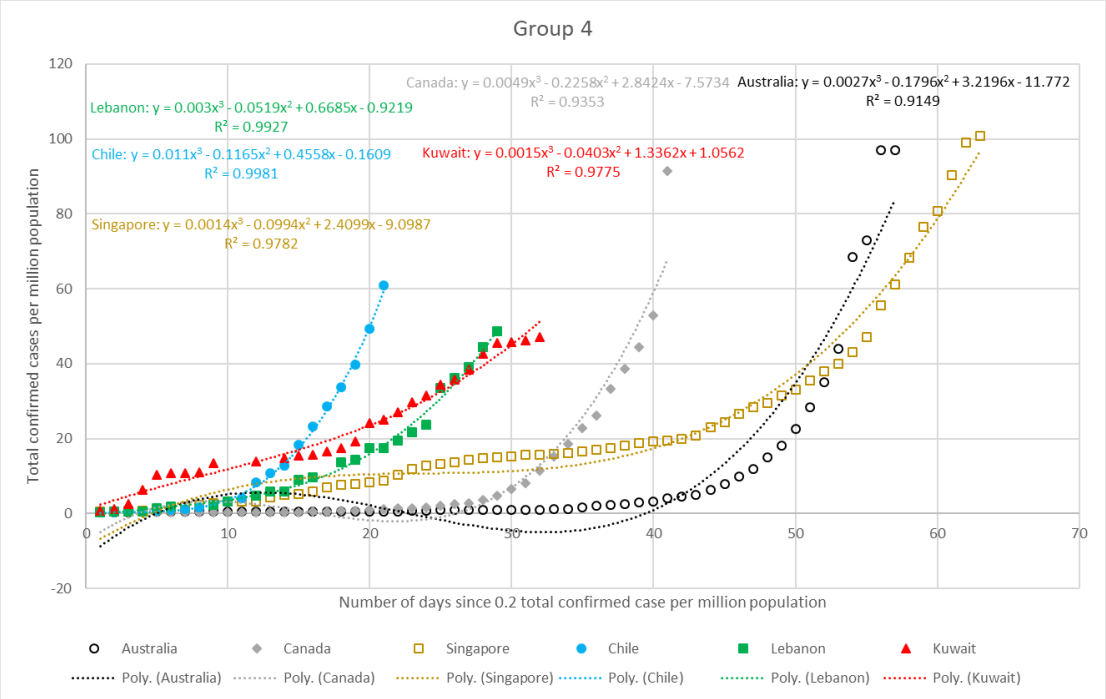


Figure 4: Group 4

3.5 Group 5

First South Korea is treated on its own, since this country presents a more complete view of the virus spread. This latter is seen to be divided into three stages according to the curve in Figure 5. The first stage here is the slow increase in the total number of cases, followed by a sharp increase in the second stage that includes an inflection point. The last stage shows the slow increase that is nearly asymptotic to a stability limit. The best fit based on the full data has an R^2 of 0.97, however when dividing the curves at the inflection point the two obtained fitted polynomials represents accurately the observed data. In fact, the percentage error obtained by the full best fit is as shown in Table 1 is 6.55% and 9.91% for 27 and 18 March respectively. These errors are reduced respectively to 0.01% and 1.05% when using the best fit of the upper curve. Furthermore, the lower curve fit estimated the following two days

from the inflection point with errors of 0.69% and 4.95%. It is important to note that this error will increase forward from the inflection point since the lower curve does not fit accurately the upper data.

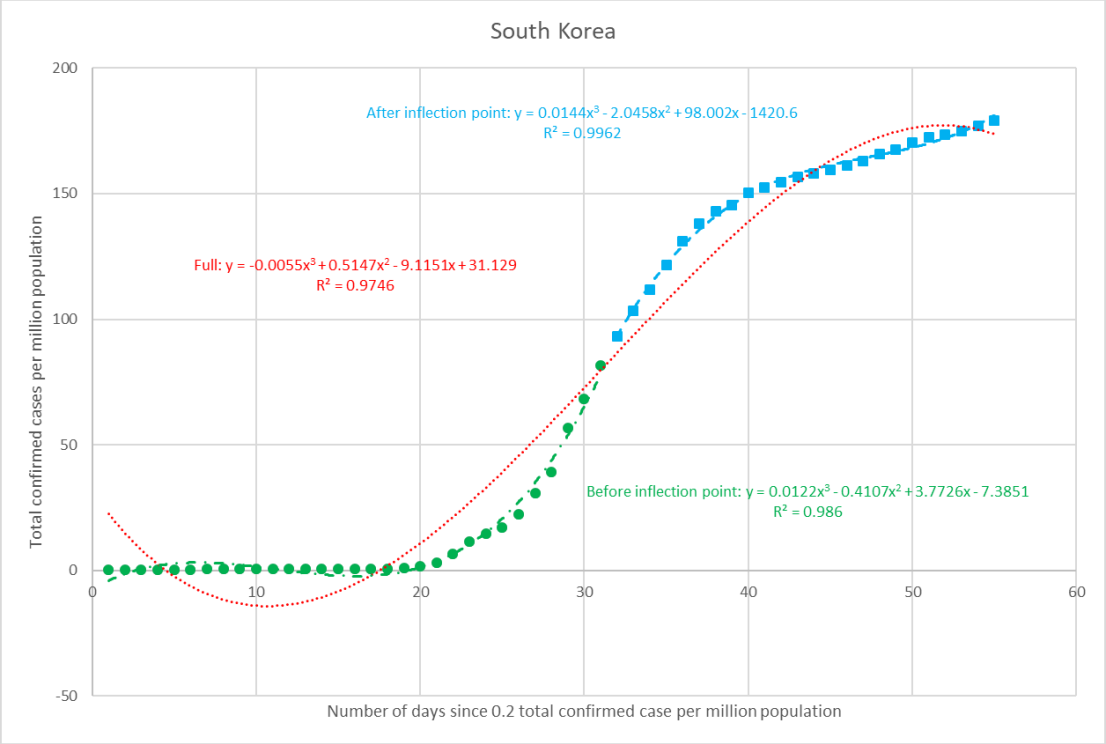


Figure 5: South Korea

The values of R^2 for the other two countries in this group are greater than 0.99 as shown in Figure 6, still small percentage errors are obtained for Greece compared to very high percentage error for Turkey as seen in Table 1. In addition, these two data were fitted with the following second degrees polynomials:

- Turkey: y = 0.278x2 - 0.4266x + 0.1521, R^2 = 0.997
- Greece: y = 0.1329x2 - 1.0808x + 2.138, R^2 = 0.9934

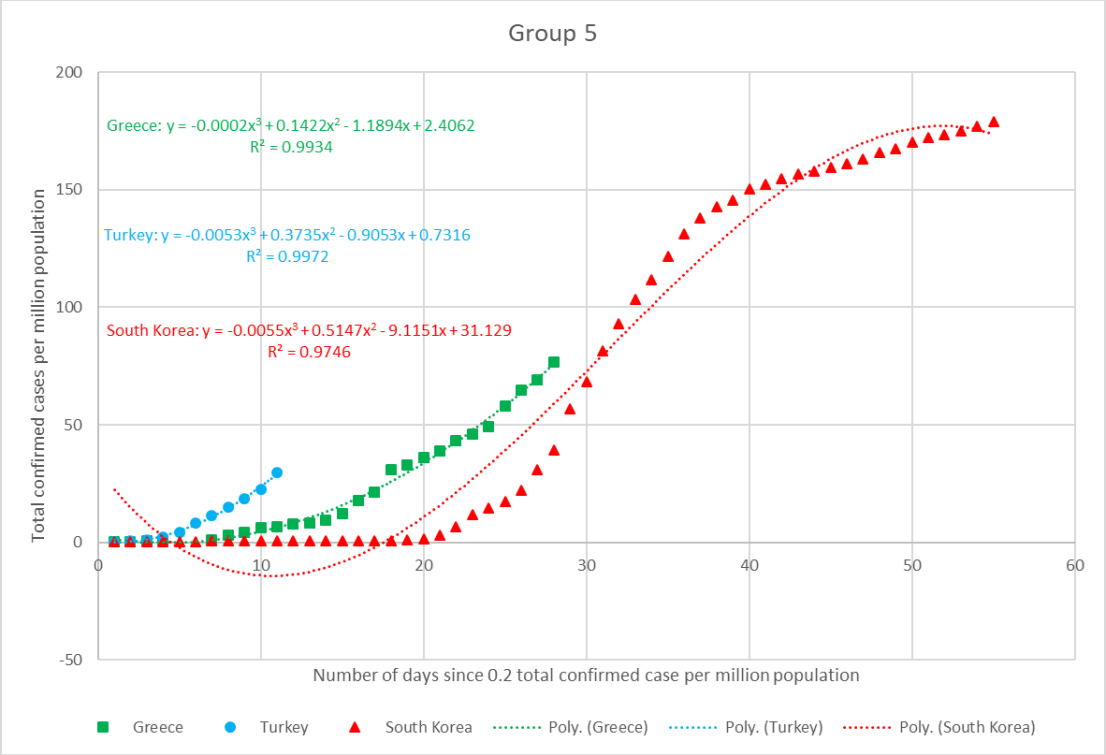
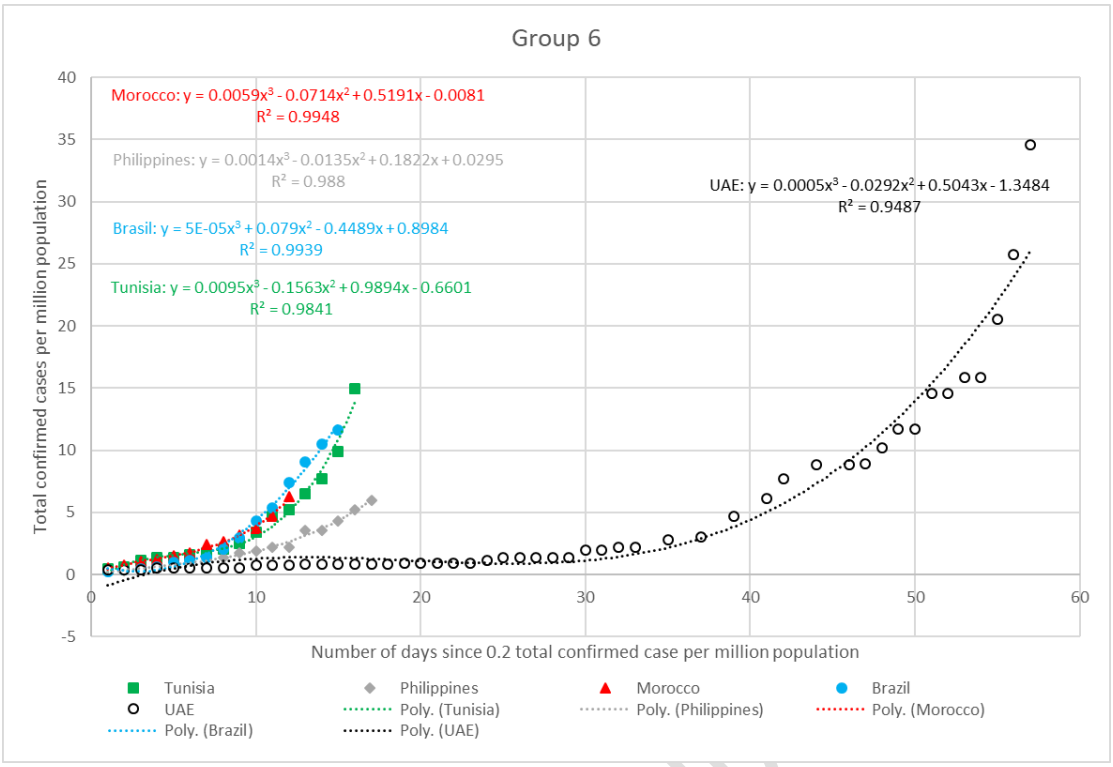


Figure 6: Group 5

3.6 Group 6

The values of R^2 are greater than 0.94. Small percentage errors are obtained for Brazil and Morocco versus high percentage error for UAE. The UAE situation is however less dangerous compared to the other groups with an average of 0.61 cases per million population over a large number of days. This is also compared with the other countries in group 6 even those who have small value of λ, that did not reach yet the sharp increase stage in the short period of time.

Since the number of days of UAE is large compared to the number of days of the remaining countries of this population, a zoomed-in version of the graph is also shown in Figure 7 bottom.



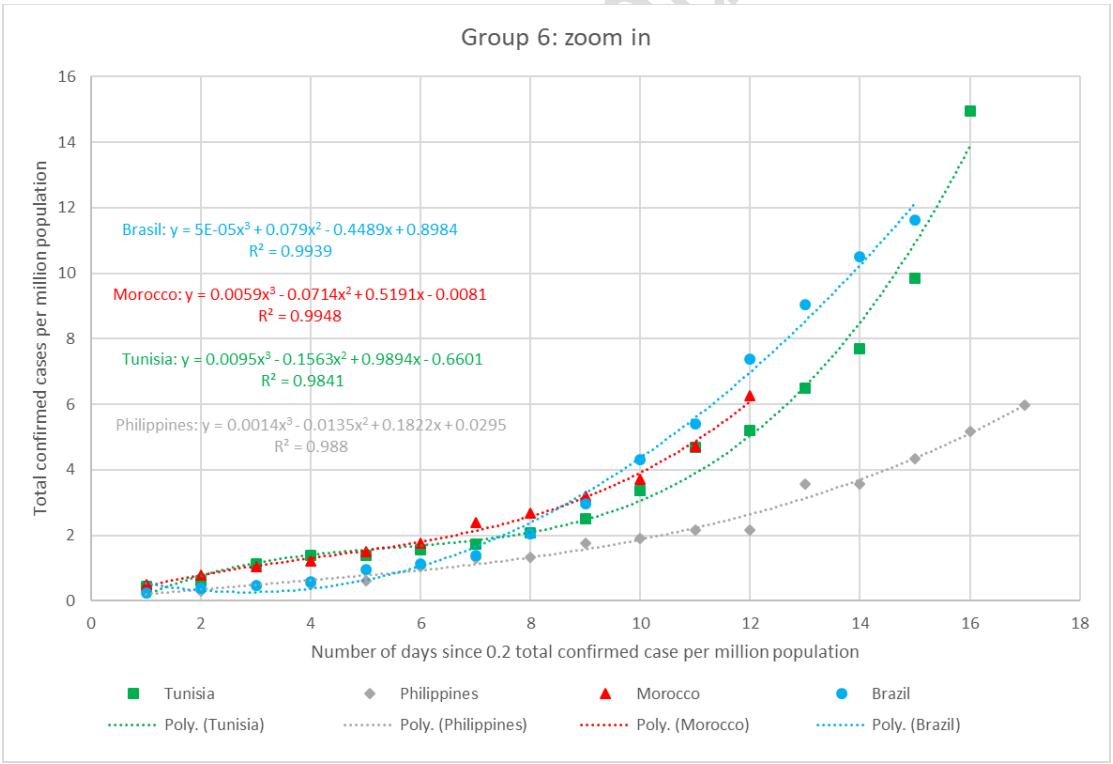


Figure 7: Group 6 Full View (Top) and Zoom-in View (bottom)

3.7 Group 7

Similar to South Korea, China represents a more complete case compared to the other countries and it also includes the three stages. However the graph shows clearly a discontinuity on February 13, the date when the Chinese government changed the way of diagnosing the virus. Based on this, the inflection point cannot be spotted but instead the curve is split on the discontinuity each part is fitted with a polynomial of degree 3 with R^2 values 0.998 for the lower curve and 0.98 for the upper curve compared to 0.9612 for the full fit.

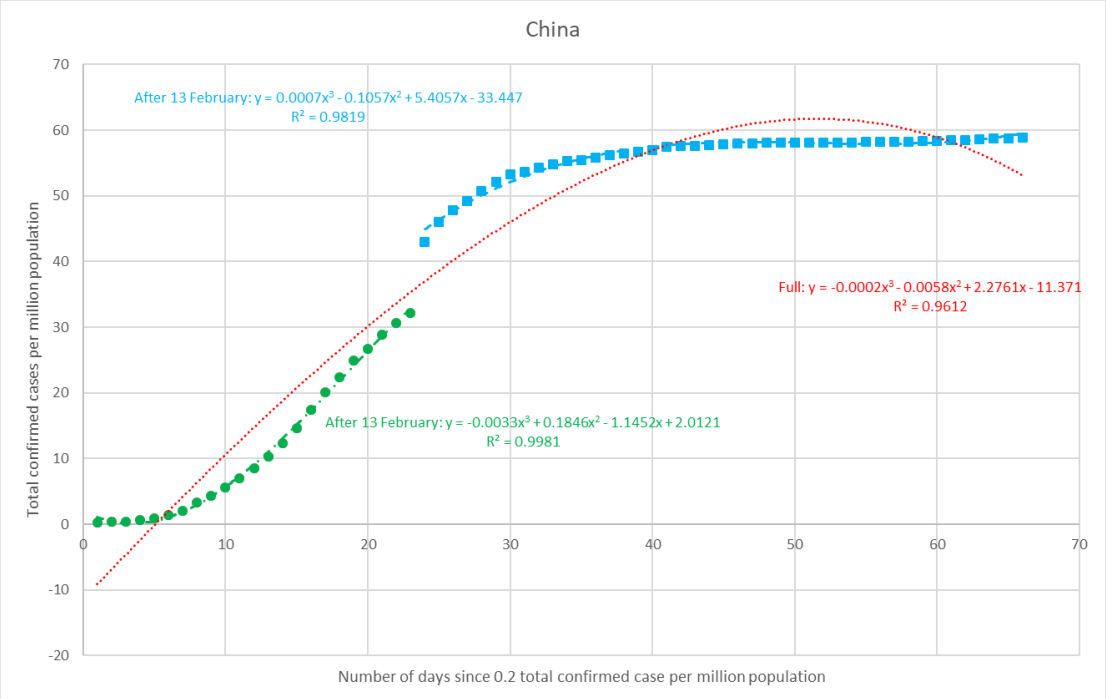


Figure 8: China

In general, for this group, small percentage errors are obtained as shown in Table 1.

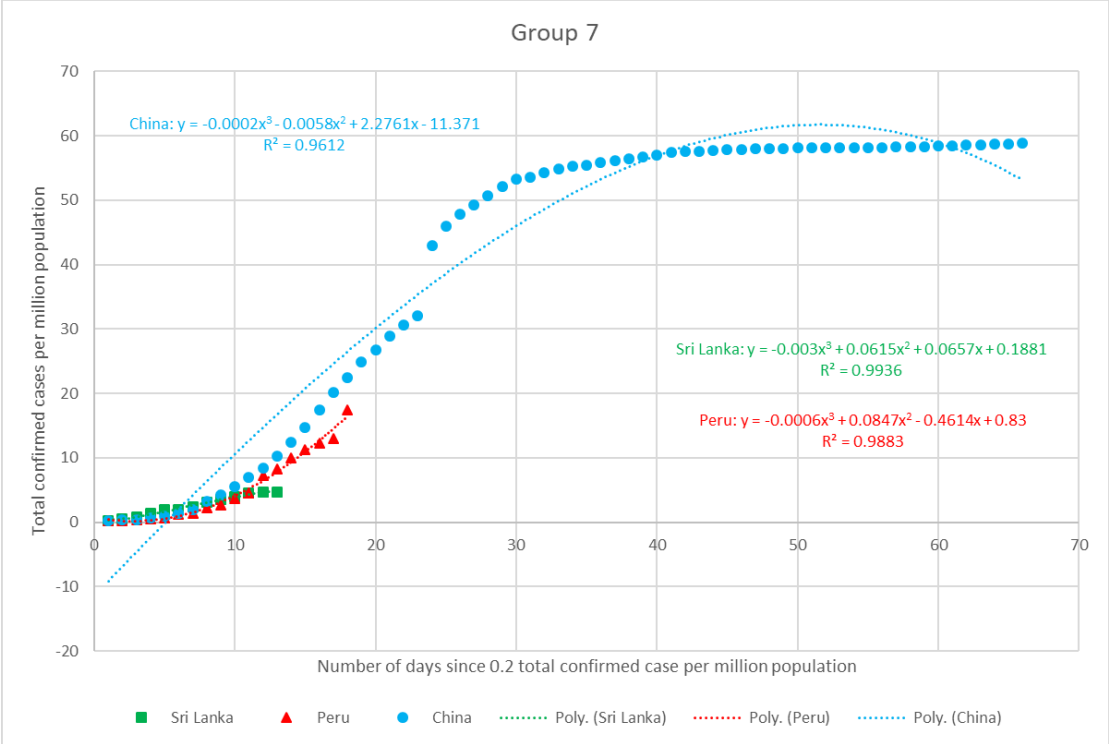


Figure 9: Group 7

4. Conclusion

Polynomials of degree 3 are good fit for the data for most of the countries in the sharp increase phase.

The countries belonging to the same categories are believed to have common behavior of the virus spread since they have comparable Poisson parameter estimate and same curvature.

The update of the data to include the following days will sure affect the coefficients of the least square polynomial and the Poisson parameter which leads to the fact that the categories are not fixed and may change with time.

The difference between South Korea and China from one side and the other countries from the other side is that the latter ones are still experiencing the sharp increase.

Based on the present categorization, if no measurements are taken, Switzerland is expected to experience a worth situation than Italy, Germany a worth situation than France, and a situation in USA similar to the current one in Germany.

References

Abdulamir, AS, & Hafidh, RR. (2020). The Possible Immunological Pathways for the Variable Immunopathogenesis of COVID—19 Infections among Healthy Adults, Elderly and Children. Electron J Gen Med. 2020; 17 (4): em202.

Deng, Yan, Liu, Wei, Liu, Kui, Fang, Yuan-Yuan, Shang, Jin, zhou, Ling, . . . Liu, Hui-Guo. (9000). Clinical characteristics of fatal and recovered cases of coronavirus disease 2019 (COVID-19) in Wuhan, China: a retrospective study. *Chinese Medical Journal, Publish Ahead of Print*. doi: 10.1097/cm9.0000000000000824

Gorbalenya, Alexander E., Baker, Susan C., Baric, Ralph S., de Groot, -Raoul J., Drosten, Christian, Gulyaeva, Anastasia A., Coronaviridae Study Group of the International Committee on Taxonomy of, Viruses. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology, 5*(4), 536-544. doi:10.1038/s41564-020-0695-z

ECDPC, European Center for Disease Prevention and Control. *Download today's data on the geographic distribution of COVID-19 cases worldwide.* Retrieved on 26 March, 2020 and 28 March, 2020, from Retrieved from https://www.ecdc.europa.eu/en/publications-

data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Tian, Sufang, Hu, Weidong, Niu, Li, Liu, Huan, Xu, Haibo, & Xiao, Shu-Yuan. Pulmonary Pathology of Early-Phase 2019 Novel Coronavirus (COVID-19) Pneumonia in Two Patients With Lung Cancer. *Journal of Thoracic Oncology*. doi: 10.1016/j.jtho.2020.02.010

WHO, World Health Organization. *A report about Coronavirus disease (COVID-19) Pandemic*. Retrieved 25 March, 2020, from Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019