The impact of vaccination on the number of COVID-19 cases

Matyáš Mattanelli & Vít Kulhánek

January 2022

Abstract.

Contents

1 Introduction

Covid-19 is one of the most frequently repeated terms in the last three years. This disease has had an effect on virtually everybody on Earth. Consequently, there are ongoing attempts to offset and eventually put an end to its negative repercussions. One possible way how to return to the "normal" living circumstances is a vaccine. It does not completely protect people against the disease, however, it is expected to reduce the quantity of hospitalizations and deaths. The aim of this project is to determine whether vaccinations also decrease the number of new cases, and thus constitute an essential tool for the elimination of the pandemic. The rest of the work is structured as follows. Section 2 provides a brief overview of the current literature, Section 3 describes the utilized data, and Section 4 is concerned with the methodology. The last two sections (5, 6) contain the discussion of the results and the conclusion, respectively.

2 Literature Review

There is only a handful of studies on similar topic since the situation is still evolving and vaccinations have only been increasing rapidly since the beginning of the second quarter of 2021. Velasco *et al.* (2021) performed a cross-country examination of which factors have a significant effect on the number of new Covid-19 cases. The authors utilized data from the beginning of the outbreak till the end of 2020 to estimate an OLS regression. Since the vaccination data were unavailable at that time, it was not included. The main findings of the study were that the number of tests and the average temperature have a significant and positive effect on the number of cases. However, we have identified some issues that may weaken the reliability of the results. Since the estimation was concerned with panel data and the individuals were represented by countries, we

expect for the unobserved heterogeneity to be present and also correlated with the independent variables. This is due to the fact that it is nearly impossible to control for all country-specific factors. For example, each country may have enacted different policies which could have a direct impact on the number of tests as well as on the number of new cases. The failure to account for the unobserved heterogeneity brings endogeneity into the model which renders OLS estimation inconsistent. We intend to improve upon this study by performing corresponding tests and employing conventional panel data estimation methods.

Toharudin *et al.* (2021) used Bayesian structural time series models including variables such as new cases, recovery cases, and number of deaths of Covid-19. They found that the vaccination program that took place in Jakarta only had a significant effect on the number of recovered cases. It is important to note that the study focused only on the very specific case of Jakarta.

Li *et al.* (2020) were testing the effects on Covid-19 in the USA. The authors found that temperature has a significant effect on the number of Covid-19 cases. Their results showed that higher temperature reduces the number of cases, however it does not affect the death rate.

To the best of our knowledge, there is still quite some unexplored area in the proposed topic. By utilizing the most recent data available, we will attempt to shed a light on the problematic whether vaccination plays a key role in the reduction of the number of Covid-19 cases.

3 Data Description

For the purpose of our analysis, we downloaded a large data set from Ritchie *et al.* (2020) which contains various information regarding Covid-19. More specifically, it includes daily observations for 207 countries. The main variables are the number of new cases of Covid-19, the number of tests determining the presence of the virus, the quantity of newly vaccinated people, total population, and many others. We aggregated the daily observations to monthly intervals in an attempt to reduce the measurement error and also to eliminate multidimensionality since most of the conventional panel data estimators are appropriate for "short" panels. As a result, we have data beginning in February 2020 and ending in December 2021.[1] Furthermore, we scaled the data by total population to provide higher cross-country comparability, thus we have variables in a form of new cases per thousand people or number of tests per thousand people.

Our additional source of the data is Weatherbase (2022). This website contains information about monthly average temperatures in over 260 countries. Based on the current literature, we consider average temperature as an important factor influencing the number of new Covid-19 cases, and therefore we include it in our model as a control variable. Since there are measurements provided for several cities in each country, we scraped the available data from the website and then computed average monthly temperatures for each country

[1] We did not include January 2022 since in the time of writing it has not ended yet.

in each month.[2]

As it was already mentioned, we are examining the effect of vaccination on the number of new cases of Covid-19. Thus, the dependent variable in our model is the number of new cases per thousand people which encompasses the quantity of positively tested people every month. After inspecting the histogram available in Figure 1, we transformed our dependent variable into logarithms because its distribution is extremely right-skewed.[3] As desired, the logarithmic transformation adjusts the distribution to be closer to Gaussian normal distribution.

Our key independent variable is the number of vaccinations per thousand people. It shows how many people were vaccinated in a particular month. This variable was also transformed into logarithms because its distribution is heavily skewed. Based on our hypothesis, we expect negative relation between the number of new cases and the number of vaccinations since the vaccine is expected to enhance individual's immunity. We define two additional control variables to ensure the robustness of our results. Li *et al.* (2020) found out that temperature is a significant factor in Covid-19 development, and therefore we include average temperature which is expected to have a negative impact on the spread of coronavirus. Lastly, the number of tests per thousand people was added as an additional control variable. The skewness of its distribution visible in the histogram forces us to use logarithms for this variable as well. We expect a positive effect of tests because without tests we cannot detect the presence of Covid-19.

The resulting data set contains quite a lot of missing observations. This is due to the fact that the first vaccination data are available in December 2020. Therefore, we disregard earlier observations. In addition, there are some missing values even after this date. We assume that they are distributed randomly and opt for listwise deletion. This is a strong assumption, however, even if it is incorrect and the missing values occur due to some country-specific characteristic, we will deal with it in the analysis. As a result, our final data set is an unbalanced panel containing 103 countries with the number of periods available ranging from 1 to 13. The descriptive statistics are presented in Table 1. We can see that in our data set there were, on average, approximately 6 monthly cases per thousand people in a country. Since the median is equal to roughly 3 cases per thousand people, the mean is likely inflated by many outliers in the upper tail. This issue should be mitigated by the logarithmic transformation since it reduces the emphasis put on extreme observations. New vaccinations and tests show a similar trend. Lastly, the average temperature in our data set is approximately 16 degrees Celsius.

Additionally, we provide the correlation matrix in Table 2 to inspect the bivariate relationship among our variables. As expected, the correlation between

[2]We utilized the Python programming language, especially the requests module. The code is available in the Appendix.

[3]We had to add a small constant to each observation since the presence of zeros renders the logarithmic transformation unfeasible. In addition, all outliers were trimmed in order to enhance the appearance of the figures

the number of new cases and the number of tests is quite high which is caused by the nature of the variables. Nobody is considered as Covid-19 positive until they receive a positive test. On the other hand, the number of vaccinations does not appear to be significantly correlated with our dependent variable. However, we expect the relationship to be more complex and thus it cannot be fully captured by a mere bivariate correlation. The direction of the relationship between new cases and average temperature matches our predictions. Regarding multicollinearity, there is not a high pairwise correlation among our independent variables so it should not cause an issue in our analysis.

4 Methodology

Since our aim is to investigate the impact of the number of newly vaccinated on the number of new Covid-19 cases using cross-country data over a certain period of time, we are essentially facing a panel data analysis. Consequently, as mentioned in the introduction, we need to consider the unobserved heterogeneity. Since we are dealing with countries, we expect them to be heterogeneous. For example, we do not control for geographical position, institutional design, culture, and many other factors that could have a direct effect on the number of new cases and which may also be correlated with our explanatory variables.[4] In addition, we expect the number of new cases to depend on its past values. If that is true, the static model could suffer from the omitted variable bias which we would like to avoid. To support our hypothesis, we test for serial correlation in a static model estimated by fixed-effects and also by first-differencing.[5] The latter should be superior in case of serially correlated errors, however, not always the serial correlation is differenced away.

The inclusion of a lagged dependent variable moves us to the area of dynamic panel data estimators. In this context, the conventional estimators such as the within estimator or the first-difference estimator are inconsistent due to the Nickell bias (Nickell 1981). One way have to restore consistency may be instrumentation. However, obtaining external instruments that are both valid and exogenous is often nearly impossible. A solution was proposed by (Arellano & Bond 1991) who devised the Difference Generalized Method of Moments estimator. The estimation procedure follows. Initially, the data is transformed using first-differencing to remove the unobserved heterogeneity. Then, with the utilization of internal instruments the resulting equation is estimated using GMM.[6] In this way, the endogeneity caused by the lagged dependent variable

[4] We perform the tests for the presence of fixed effects only in the static model since they rely either on the residuals of the pooled model (LM test) or on the comparison of within and pooled OLS estimators (F test). Since both are inconsistent in the dynamic panel data analysis framework, we have to rely on our intuition. The same applies to the Hausman test.

[5] We perform the Wooldridge Test for AR(1) Errors in FE Panel Models and the Wooldridge first–difference–based test for AR(1) errors in first–differenced panel models.

[6] The instruments are the lagged levels of the included variables. These instruments are exogenous under the assumption of no serial correlation. The moment conditions are then derived based on the instrument exogeneity assumption.

and also any other endogenous variables is dealt with. This is very useful in our case since our variables may be subjected to a measurement error. As Roodman (2009a) suggests, the Difference GMM estimator is appropriate for "short" panels with many individuals, models with possibly endogenous regressors, and the presence of fixed-effects which is precisely our case. In addition, the two-step GMM procedure ensures robustness to heteroskedasticity.

We also need to consider that in case that our dependent variable is highly persistent, the Difference GMM estimator suffers from the weak instruments problem. A solution was proposed by Arellano & Bover (1995) and Blundell & Bond (1998) who developed the System Generalized Method of Moments estimator. The authors build on the Difference GMM by introducing additional moment conditions. In addition to the differenced equation instrumented by lagged levels of the variables, a level equation instrumented by lagged first differences is estimated. These further moment conditions should solve the weak instruments issue.

To verify the reliability of the results, we perform the Hansen-Sargan test and the Arellano-Bond test which are essential for the validity of internal instruments. The former is a test of overidentifying restrictions while the latter tests for serial correlation in the differenced errors. The order of the serial correlation depends on the number of included lags. If two lags are included, the serial correlation of third order is tested. Therefore we choose $(p-1)$ lags so the null hypothesis of no serial correlation of the p-th order cannot be rejected. We also make sure to restrict the number of instruments to avoid instrument proliferation as indicated by Roodman (2009b). This pertains to ensuring that the p-value of the Hansen-Sargan test is not too high. In addition, the author offers a simple rule of thumb that the number of instruments should not exceed the number of individuals. Lastly, we use the Hausman test to ascertain whether our variables are indeed subjected to the measurement error.

As a result, we estimate the following model:

$$\begin{aligned}\log(New\_cases_{it}) =& \beta_1 \log(New\_cases_{i(t-1)}) + \beta_2 \log(New\_cases_{i(t-2)}) + \\ &+ \beta_3 \log(New\_vaccinations_{i(t-1)}) + \\ &+ \beta_4 \log(New\_vaccinations_{i(t-2)}) + \beta_5 \log(New\_tests_{it}) + \\ &+ \beta_6 Average\_temperature_{it} + Month_t + a_i + \epsilon_{it}\end{aligned}$$

where $Month_t$ is a vector of dummy variables for each month capturing the time effects, a_i is the time-invariant unobserved heterogeneity, and ϵ_{it} is the idiosyncratic error. We expect new vaccinations to affect the number of new cases with a lag since it takes some time for the vaccine to become effective. On the other hand, we anticipate the number of tests and the average temperature to only have a contemporary effect.

5 Empirical results

Firstly, we provide the results of the specification tests in Table 3. The first two tests test the null hypothesis of no serial correlation in the errors in the static

model estimated by the within estimator and the first-differenced estimator, respectively.[7] Both tests firmly reject the null hypothesis at any conventional significance level and thus support our choice of a dynamic model. The Breusch-Pagan test rejects the null hypothesis of homoskedasticity at the 5% significance level which indicates that we should opt for the two-step GMM estimation to make our results robust to heteroskedasticity. We also perform three additional tests on the static model, however, the results are merely indicative. Both the LM test and the F test reject the null hypothosis of no significant individual and time effects. In addition, the Hausman test rejects the null hypothesis of both the FGLS estimator and the within estimator being consistent in favor of the alternative that only the within estimator is consistent. As a result, we have some evidence for the presence of the unobserved heterogeneity in the static model which is also correlated with our independent variables. Therefore, we will estimate the dynamic model with the Difference and System GMM estimators. The last two specification tests test for the second-order serial correlation in the dynamic model with one lag of the dependent variable. Irrespective of the estimation method, the null hypothesis of no serial correlation is rejected. Consequently, we will include a second lag of the dependent variable into our model as well.

The estimation results are available in Table 4. As explained above, we utilize two estimators (Difference GMM and System GMM) to estimate the model with different specifications. Firstly, it is essential to verify the exogeneity of our instruments. In all cases, the p-value of the Hansen-Sargan test is sufficiently high not to reject the null hypothesis of instrument exogeneity but also adequately low to avoid instrument proliferation. The Arellano-Bond test does not reject the null hypothesis of serial correlation in the errors in any specification which further supports the exogenity of the instruments and also shows that we included the correct number of lags. With respect to the weak instruments problem, it appears that our dependent variable is highly persistent, and therefore the Difference GMM estimator underestimates the true value of the coefficient of the lagged dependent variable. As a result, we will rely mostly on the System GMM estimator.

Secondly, we utilized the Hausman test in order to choose the correct specification of the model. In columns (1) and (2) we instrument only the lagged dependent variable and thus assume that other variables are exogenous. However, since we suspect that the number of new cases and the number of new vaccinations may be subjected to the measurement error, we instrument them in columns (3) and (4). Then we compare both specification using the Hausman test under the null hypothesis that both models are consistent, and therefore by instrumenting we are merely losing efficiency. However, since irrespective of the estimation method the low p-value of the test leads to the rejection of the null hypothesis, we conclude that there appears to be an endogeneity problem which requires instrumentation. Therefore, we regard the models in columns (3) and

[7] "Static" refers to the model defined above, however, the lagged dependent variables are not included

(4) as our baseline results. In column (5) we estimate a model with only one lag of the dependent variable by Difference GMM since after instrumentation the second lag became insignificant. In this way we test the sensitivity of our results.

Lastly, we provide the interpretation of our results. Referring to column (4) which presents our most reliable estimates, we can see that both lags of our variable of interest are significant and appear to have a negative effect on the number of Covid-19 cases. More specifically, the results indicate that a one percentage increment in the number of vaccinations per thousand members of the population in the previous month decreases the number of new cases per thousand members of the population in the current month by approximately 0.09%. The effect appears to be even stronger after two months when the decrease is approximately 0.11%. However, the coefficient of the second lag is significant only in two specifications and when the Difference GMM estimator is used, it is positive. Therefore, we need to treat it with caution. On the other hand, the sign of the coefficient of the first lag is the same across specifications and it is also significant in almost all cases. Consequently, based on our analysis, the number of vaccinations appears to negatively effect the number of cases with a one month lag. With regard to our control variables, the number of tests appears to have a positive and significant effect in all specifications. The magnitude of its coefficient is considerable as well. It suggests that a one percentage increase in the number of tests per thousand members of the population results in approximately a 0.34% increase in the number of cases. Average temperature is insignificant and the sign of its coefficient differs across specifications. This may be likely due to the inclusion of the lags of the dependent variable since they explain a lot of variation. However, our variables of interest is robust to their inclusion which signals a strong effect.

6 Conclusion

The aim of this project was to analyze the effect of the number of new vaccinations on the number of new Covid-19 cases. For this purpose we utilized a large cross-country data set with monthly observations. Given the inclusion of lags of the dependent variable, we performed a dynamic panel data analysis. With the adoption of the Difference and System GMM estimators, we found that vaccinations appear to have a significant and negative effect on the number of new cases with a one month lag. Our results are robust to the inclusion of various control variables. Nevertheless, we are aware that our analysis is not without limitations. Firstly, by monthly aggregation we shifted our focus from the time dimension to the cross-sectional dimension which is common practice in panel data analysis. However, it is possible to approach the issue in a more time series oriented way which presents and opportunity for future research. Secondly, the presence of vaccinations is still quite recent and in some countries it is not that well documented. Therefore, it is difficult to draw any firm conclusions. Thirdly, we were not able to control for mutations of the coronavirus since the

first vaccination data are available only after the first emergence of the delta mutation and it is too soon to control for the omicron variant. This may be a slight setback since different mutations may respond to vaccinations differently. Similarly, we cannot control for vaccination types since our observations are on the country level. Despite the limitations, however, we believe that our analysis supports the importance of vaccinations in the battle against the Covid-19 pandemic.

Bibliography

Arellano, M. & S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* **58(2)**: pp. 277–297.

Arellano, M. & O. Bover (1995): "Another look at the instrumental variable estimation of error-components models." *Journal of Econometrics* **68(1)**: pp. 29–51.

Blundell, R. & S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics* **87(1)**: pp. 115–143.

Li, A. Y., T. C. Hannah, J. R. Durbin, N. Dreher, F. M. McAuley, N. F. Marayati, Z. Spiera, M. Ali, A. Gometz, J. Kostman, & T. F. Choudhri (2020): "Multivariate analysis of factors affecting covid-19 case and death rate in u.s. counties: The significant effects of black race and temperature." *medRxiv* .

Nickell, S. (1981): "Biases in dynamic models with fixed effects." *Econometrica* **49(6)**: pp. 1417–26.

Ritchie, H., E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, & M. Roser (2020): "Coronavirus pandemic (covid-19)." *Our World in Data* Available at `https://ourworldindata.org/coronavirus`. [Accessed 2022-01-16].

Roodman, D. (2009a): "How to do xtabond2: An introduction to difference and system gmm in stata." *The Stata Journal* **9(1)**: pp. 86–136.

Roodman, D. (2009b): "A note on the theme of too many instruments*." *Oxford Bulletin of Economics and Statistics* **71(1)**: pp. 135–158.

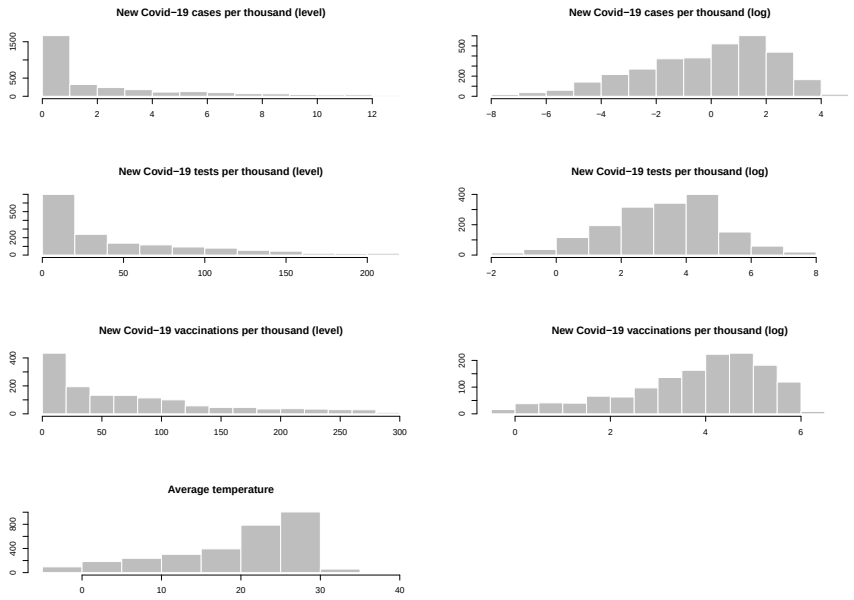
Toharudin, T., R. S. Pontoh, R. E. Caraka, S. Zahroh, P. Kendogo, N. Sijabat, M. D. P. Sari, P. U. Gio, M. Basyuni, & B. Pardamean (2021): "National vaccination and local intervention impacts on covid-19 cases." *Sustainability* **13(15)**.

Velasco, J. M., W.-C. Tseng, & C.-L. Chang (2021): "Factors affecting the cases and deaths of COVID-19 victims." *Int J Environ Res Public Health* **18(2)**.

Weatherbase (2022): "Travel weather averages (weatherbase)." Available at `https://www.weatherbase.com/`. [Accessed 2021-01-16].

Appendix

Figure 1: Histograms



Source: Authors' computations based on the compiled data set

Table 1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Median	Max
New cases	5.753	7.642	0.003	2.978	63.576
New vaccinations	97.935	95.244	0.000	71.717	570.621
New tests	126.222	269.659	0.032	42.274	3,089.120
Average temperature	16.443	9.320	−17.858	17.354	34.660

Note: All variables apart from Average Temperature are in per thousand people terms

Table 2: Correlation matrix

	New cases	New tests	New vaccinations	Average temperature
New cases	1			
New tests	0.63*	1		
New vaccinations	−0.00027	0.019	1	
Average temperature	−0.32*	−0.39*	−0.016	1

Note: All variables apart from Average temperature are in logs per thousand people
* Significant at 95%

Table 3: Specification tests

Wooldridge test (FE)*	0.00
Wooldridge test (FD)*	0.00
Breusch-Pagan test*	0.01
LM test*	0.00
F test*	0.00
Hausman test*	0.00
Arellano-Bond test (D-GMM)**	0.09
Arellano-Bond test (Sys-GMM)**	0.00

* All the tests were performed on the static model
** Tests the second order serial correlation in the dynamic model with only one lag of the dependent variable included

Table 4: Estimation results

	Dependent variable:				
	$\log(New\_cases_{it})$				
	D-GMM	SYS-GMM	D-GMM	SYS-GMM	D-GMM
	(1)	(2)	(3)	(4)	(5)
$\log(New\_cases_{i(t-1)})$	0.750^{***}	1.013^{***}	0.331^{*}	1.013^{***}	0.224^{**}
	(0.166)	(0.061)	(0.170)	(0.068)	(0.101)
$\log(New\_cases_{i(t-2)})$	-0.343^{***}	-0.534^{***}	0.038	-0.452^{***}	
	(0.113)	(0.054)	(0.190)	(0.061)	
$\log(New\_vaccinations_{i(t-1)})$	-0.110^{**}	-0.006	-0.292^{**}	-0.092^{**}	-0.383^{***}
	(0.054)	(0.031)	(0.142)	(0.039)	(0.146)
$\log(New\_vaccinations_{i(t-2)})$	-0.050	-0.041	0.147^{**}	-0.113^{***}	0.090
	(0.033)	(0.025)	(0.062)	(0.043)	(0.064)
$\log(New\_tests_{it})$	0.851^{***}	0.246^{***}	1.653^{***}	0.344^{***}	2.444^{***}
	(0.284)	(0.050)	(0.444)	(0.091)	(0.358)
$Average\_temperature_{it}$	-0.012	-0.010	-0.004	0.004	0.012
	(0.021)	(0.008)	(0.017)	(0.009)	(0.026)
No. of countries	103	103	103	103	103
No. of instruments	69	75	41	76	60
Hansen-Sargan test (p-value)	0.13	0.13	0.29	0.15	0.22
Arellano-Bond test (p-value)	0.74	0.42	0.29	0.44	0.98
Hausman test (p-value)	-	-	0.00	0.00	-

*p<0.1; **p<0.05; ***p<0.01

Odd columns (1), (3), and (5) are estimated by the Difference GMM estimator while the even columns (2) and (4) are estimated using System GMM. In columns (1) and (2) only the lagged dependent variables are instrumented. In rest of the columns, all variables apart from $Average\_temperature_{it}$ are instrumented. The p-value of the Hausman test in column (3) is based on the comparison of models from columns (1) and (3). Similarly, the p-value in column (4) is based on the comparison of models from columns (2) and (4). All variables apart from $Average\_temperature_{it}$ are in per thousand members of the population terms.

R code

```
############################
### Data pre-processing ###
############################

#Importing the covid data set
library(readr)
owid_covid_data <- read_csv("owid-covid-data.csv")
owid_covid_data<-as.data.frame(owid_covid_data)

#Extracting relevant columns
dataset <- owid_covid_data[,c(3,4,6,26,39,49)]

#Filtering out non-countries
non_countries <- c("Africa","Asia","European Union","Europe",
                   "High income","International",
                   "Lower middle income","Low income",
                   "North America","Soouth Africa",
                   "South America","Upper middle income",
                   "World")
dataset_filtered <- dataset[!dataset$location %in%
                              non_countries,]

### Aggregating daily data to monthly data ###

#Writing a function to sum properly (we want NA when
#all arguments are NA, not a 0)
proper_sum <- function(x){
  if (all(is.na(x))){
    value <- NA
  } else {
    value <- sum(x,na.rm=T)
  }
  return(value)
}

#Extracting the month and year for aggregation
dataset_filtered$month<-months(dataset_filtered$date)
dataset_filtered$year<-as.numeric(format(dataset_filtered$date,
                                         "%Y"))

#Aggregation using the dplyr package
library(dplyr)
dataset_monthly <- dataset_filtered %>%
  group_by(location,month,year) %>%
```

```
  summarise(new_cases=proper_sum(new_cases),
            new_tests=proper_sum(new_tests),
            new_vaccinations=proper_sum(new_vaccinations),
            population=unique(population))

#Adding back the date (time dimension)
dataset_monthly$date<-as.Date(paste("01",
                                    dataset_monthly$month,
                                    dataset_monthly$year),
                              "%d_%B_%Y")

#Adjusting the data set
dataset_semifinal<-dataset_monthly[,c(1,8,4:7)]
dataset_semifinal<-dataset_semifinal[order(dataset_semifinal$location,
                                           dataset_semifinal$date),]

#Loading average temperature
avg_temp <- read_csv('avg_temp.csv')

#Converting months to dates
avg_temp$date <- as.Date(paste("01",avg_temp$Month,
                               "2020"), "%d_%B_%Y")

#Adding the year 2021
avg_temp_2021 <- avg_temp
avg_temp_2021$date <- as.Date(paste("01",avg_temp_2021$Month,
                                    "2021"), "%d_%B_%Y")

#Concatenating the two data sets
avg_temp_semifinal <- rbind(avg_temp,avg_temp_2021)

#Reordering the new data set
avg_temp_semifinal<-avg_temp_semifinal[order(avg_temp_semifinal$Country,
                                             avg_temp_semifinal$date),]

#Finalizing the data set
avg_temp_final <- avg_temp_semifinal[,c(1,4,3)]
colnames(avg_temp_final)[c(1,3)] <- c("location","avg_temp")

### Merging avg_temp with the main data set ###

#Looking for dissimilar country names
setdiff(unique(dataset_semifinal$location),
        unique(avg_temp_final$location))
setdiff(unique(avg_temp_final$location),
        unique(dataset_semifinal$location))
```

```
#Unifying the country names
avg_temp_final$location<-gsub("Democratic Republic of the Congo",
                              "Democratic Republic of Congo",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Czech Republic","Czechia",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("United States of America",
                              "United States",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Federated States of Micronesia",
                              "Micronesia (country)",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("East Timor","Timor",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Faroe Islands","Faeroe Islands",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Macedonia","North Macedonia",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("The Gambia","Gambia",
                              avg_temp_final$location)
avg_temp_final$location[avg_temp_final$location=="Republic of Congo"]="Congo"
avg_temp_final$location<-gsub("Pitcairn Islands","Pitcairn",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Cura ao","Curacao",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("The Bahamas","Bahamas",
                              avg_temp_final$location)
avg_temp_final$location<-gsub("Sint Maarten","Sint Maarten (Dutch part)",
                              avg_temp_final$location)
#Adding data for "Jersey" and "Guernsey"
jersey=data.frame(location=rep("Jersey",24),
                  date=seq.Date(as.Date("2020-01-01"),
                                as.Date("2021-12-01"),by="month"),
                  avg_temp=rep(c(6.3,6.1,7.9,9.5,12.6,15.1,17.2,17.5,
                                 15.8,13,9.6,7.1),2))
guernsey=data.frame(location=rep("Guernsey",24),
                    date=seq.Date(as.Date("2020-01-01"),
                                  as.Date("2021-12-01"),by="month"),
                    avg_temp=rep(c(6.9,6.5,7.8,9.3,12.1,14.6,16.6,17,
                                   15.6,13,10,7.8),2))
avg_temp_final2<-rbind(avg_temp_final,jersey,guernsey)
avg_temp_final2<-avg_temp_final2[order(avg_temp_final2$location,
                                       avg_temp_final2$date),]
#Merging
dataset_semifinal2<-merge(dataset_semifinal,avg_temp_final2,
                          by=c("location","date"),all.x = T)
```

```
#Exporting the data set
write_csv(dataset_semifinal2,"dataset.csv")

#Importing the data set
library(readr)
dataset <- read_csv("dataset.csv")
dataset <- as.data.frame(dataset)

########################
### Data inspection ###
########################

#Inspecting the data set
summary(dataset)
#Negative cases + Vaccination many missing values

dataset$new_cases[dataset$new_cases<0]<-NA
#Negative cases do not make sense

min(na.omit(dataset[,c(1:5,9)])$date)
#Earliest vaccination data in December 2020

dataset<-dataset[dataset$date>=as.Date("2020-10-01"),]
#We retain some past values given the inclusion of lags in the analysis

#Scaling by population (for increased cross-country comparability)
dataset$new_cases_per_thousand<-dataset$new_cases/(dataset$population/1000)
dataset$new_vaccinations_per_thousand<-
  dataset$new_vaccinations/(dataset$population/1000)
dataset$new_tests_per_thousand<-dataset$new_tests/(dataset$population/1000)

#Histograms
trim <- function(x){#Defining a trimming function to make the histograms nicer
  x[(x > quantile(x, 0.25,na.rm=T)-1.5*IQR(x,na.rm=T)) &
      (x < quantile(x, 0.75,na.rm=T)+1.5*IQR(x,na.rm=T))]
}

par(mfrow=c(4,2))
hist(trim(dataset$new_cases_per_thousand),border = F,col="grey",
     main = "New Covid-19 cases per thousand (level)",xlab = "",
     ylab="") #Extremely skewed, many outliers
hist(trim(log(dataset$new_cases_per_thousand+
  min(dataset$new_cases_per_thousand[dataset$new_cases_per_thousand>0],
                      na.rm = T))),border=F,col="grey",xlab="",ylab="",
     main="New Covid-19 cases per thousand (log)")
```

```
#A bit better, need to add a constant due to zeros

hist(trim(dataset$new_tests_per_thousand),border = F,col="grey",
     main = "New Covid-19 tests per thousand (level)",xlab = "",
     ylab="") #Heavily skewed
hist(trim(log(dataset$new_tests_per_thousand)),border=F,col="grey",
     xlab="",ylab="",main="New Covid-19 tests per thousand (log)")

hist(trim(dataset$new_vaccinations_per_thousand),border = F,col="grey",
     main = "New Covid-19 vaccinations per thousand (level)",xlab = "",
     ylab="") #Skeeeewed
hist(trim(log(dataset$new_vaccinations_per_thousand+
min(dataset$new_vaccinations_per_thousand[dataset$new_vaccinations_per_thousand>
                                          0],na.rm = T))),border=F,
col="grey",xlab="",ylab="",
main="New Covid-19 vaccinations per thousand (log)") #A lot better

hist(trim(dataset$avg_temp),border=F,col="grey",
     main = "Average temperature",xlab = "",
     ylab="") #A bit left-skewed (but log won't help)

#Transforming the variables
dataset$log_new_cases_per_thousand <-
  log(dataset$new_cases_per_thousand+
        min(dataset$new_cases_per_thousand[dataset$new_cases_per_thousand>0],
            na.rm = T))

dataset$log_new_tests_per_thousand <- log(dataset$new_tests_per_thousand)
dataset$log_new_vaccinations_per_thousand <-
log(dataset$new_vaccinations+
min(dataset$new_vaccinations_per_thousand[dataset$new_vaccinations_per_thousand
                                          >0],na.rm = T))

#Descriptive statistics
library(stargazer)
stargazer(na.omit(dataset[,c(10:12,9)]),title = "Descriptive statistics",
          label="tab:des",header = F,
          summary.stat = c("mean","sd","min","median","max"),
notes="Note: All variables apart from Average Temperature are in per
thousand people terms")
length(unique(na.omit(dataset)$location)) #103 countries
table(table(na.omit(dataset)$location)) #1 to 13 periods

#Correlation matrix
library(rstatix)
cor_mat<-cor_mat(dataset[,c(13:15,9)])
```

```
cor_mat<-cor_mark_significant(cor_mat, cutpoints = c(0,0.05,1),
                                symbols = c("*",""))
for (i in 1:4){
  cor_mat[i,i+1]<-1
}
stargazer(cor_mat,summary = F, header = F, title = "Correlation matrix",
notes = c("Note: All variables apart from Average temperature are in
          logs per thousand people","* significant at 95%"),
          label="tab:cor_mat",rownames = F)

################
### Analysis ###
################

### Static FE ###

library(plm)
model_fe_static <- plm(log_new_cases_per_thousand~
                         lag(log_new_vaccinations_per_thousand,
                         1:2)+log_new_tests_per_thousand+avg_temp,
                  data=dataset,model="within",effect = "twoways",
                  index=c("location","date"))
pwartest(model_fe_static) #strong evidence of serial correlation
pwfdtest(log_new_cases_per_thousand~log_new_tests_per_thousand+
           lag(log_new_vaccinations_per_thousand,1:2)+avg_temp,
         data=dataset,index=c("location","date"))
#serial correlation in differenced errors as well

library(lmtest)
bptest(model_fe_static) #evidence of heteroskedasticity
plmtest(log_new_cases_per_thousand~lag(log_new_vaccinations_per_thousand,1:2)
        +log_new_tests_per_thousand+avg_temp,
        data=dataset,effect = "twoways",index=c("location","date"))
#Significant effects

pFtest(log_new_cases_per_thousand~lag(log_new_vaccinations_per_thousand,1:2)
       +log_new_tests_per_thousand+avg_temp,
       data=dataset,effect = "twoways",index=c("location","date"))
#Significant effects

phtest(log_new_cases_per_thousand~lag(log_new_vaccinations_per_thousand,1:2)
       +log_new_tests_per_thousand+avg_temp,
       data=dataset,effect = "twoways",index=c("location","date"),
       model = c("within", "random"))

### Difference GMM ###
```

```
#One lag Difference GMM
diff_GMM_model_one_lag<-pgmm(log_new_cases_per_thousand~
                             lag(log_new_cases_per_thousand,1)+
                             lag(log_new_vaccinations_per_thousand,1:2)+
                             log_new_tests_per_thousand+avg_temp|
                             lag(log_new_cases_per_thousand,2:7),
                           data=na.omit(dataset),effect = "twoways",
                           model=c("twosteps"),transformation = "d",
                           index=c("location","date"))
summary(diff_GMM_model_one_lag)
#Arellano-Bond test rejects the null hypothesis of no 2nd order serial
#correlation at the 10% significance level => We need two lags

#Two lags Difference GMM
diff_GMM_model_two_lags <- pgmm(log_new_cases_per_thousand~
                                  lag(log_new_cases_per_thousand,1:2)+
                                lag(log_new_vaccinations_per_thousand,1:2)+
                                log_new_tests_per_thousand+avg_temp|
                                  lag(log_new_cases_per_thousand,3:99),
                           data=na.omit(dataset),effect = "twoways",
                           model=c("twosteps"),transformation = "d",
                           index=c("location","date"))
summary(diff_GMM_model_two_lags)
#Sargan p-value 0.13 => instruments valid, no proliferation

mtest(diff_GMM_model_two_lags,3)
#p-value 0.74 => No serial correlation of 3rd order => instruments valid

sargan_test_diff_model_two_lags<-sargan(diff_GMM_model_two_lags)
as.numeric(sargan_test_diff_model_two_lags$parameter)+
  length(coefficients(diff_GMM_model_two_lags)) #69 instruments

#Instrumenting new tests and vaccination to account for the measurement error
diff_GMM_model_two_lags_endo<-pgmm(log_new_cases_per_thousand~
                                   lag(log_new_cases_per_thousand,1:2)+
                                   lag(log_new_vaccinations_per_thousand,1:2)+
                                   log_new_tests_per_thousand+avg_temp|
                                   lag(log_new_cases_per_thousand,3)+
                                   lag(log_new_tests_per_thousand,2)+
                                 lag(log_new_vaccinations_per_thousand,3),
                                  data=na.omit(dataset),effect = "twoways",
                                    model=c("twosteps"),transformation = "d",
                                          index=c("location","date"))
summary(diff_GMM_model_two_lags_endo)
#Sargan p-value 0.29 => instruments valid, slight proliferation
```

```
mtest(diff_GMM_model_two_lags_endo,3)
#p-value 0.29 => No serial correlation of 3rd order => instruments valid

sargan_test_diff_model_two_lags_endo<-sargan(diff_GMM_model_two_lags_endo)
as.numeric(sargan_test_diff_model_two_lags_endo$parameter)+
  length(coefficients(diff_GMM_model_two_lags_endo)) #41 instruments
phtest(diff_GMM_model_two_lags, diff_GMM_model_two_lags_endo)
#Null rejected => Endogeneity was likely present

#Including only one lag and instrumenting the endogenous variables
diff_GMM_model_one_lag_endo <- pgmm(log_new_cases_per_thousand~
                                    lag(log_new_cases_per_thousand,1)+
                                    lag(log_new_vaccinations_per_thousand,1:2)+
                                    log_new_tests_per_thousand+avg_temp|
                                    lag(log_new_cases_per_thousand,2:4)+
                                    lag(log_new_tests_per_thousand,2)+
                                    lag(log_new_vaccinations_per_thousand,3),
                                    data=na.omit(dataset),effect = "twoways",
                                    model=c("twosteps"),transformation = "d",
                                        index=c("location","date"))
summary(diff_GMM_model_one_lag_endo)
#Sargan p-value 0.22 => instruments valid,
#no proliferation, no serial correlation of 2nd order (p-value 0.98)
sargan_test_diff_model_one_lag_endo<-sargan(diff_GMM_model_one_lag_endo)
as.numeric(sargan_test_diff_model_one_lag_endo$parameter)+
  length(coefficients(diff_GMM_model_one_lag_endo)) #60 instruments

### System GMM ###

#One lag System GMM
system_GMM_model_one_lag <- pgmm(log_new_cases_per_thousand~
                                 lag(log_new_cases_per_thousand,1)+
                                 lag(log_new_vaccinations_per_thousand,1:2)+
                                 log_new_tests_per_thousand+avg_temp|
                                 lag(log_new_cases_per_thousand,2:7),
                                 data=na.omit(dataset),effect = "twoways",
                                 model=c("twosteps"),transformation = "ld",
                                     index=c("location","date"))
summary(system_GMM_model_one_lag)
#Arellano-Bond test rejects the null hypothesis of no
#2nd order serial correlation => We need two lags

#Two lags System GMM
system_GMM_model_two_lags <- pgmm(log_new_cases_per_thousand~
                                  lag(log_new_cases_per_thousand,1:2)+
```

```
                                    lag(log_new_vaccinations_per_thousand,1:2)+
                                    log_new_tests_per_thousand+avg_temp|
                                    lag(log_new_cases_per_thousand,3:8),
                                    data=na.omit(dataset),effect = "twoways",
                                    model=c("twosteps"),transformation = "ld",
                                      index=c("location","date"))
summary(system_GMM_model_two_lags)
#Sargan p-value 0.13 => instruments valid, no proliferation

mtest(system_GMM_model_two_lags,3)
#p-value 0.42 => No serial correlation of 3rd order => instruments valid

sargan_test_system_model_two_lags<-sargan(system_GMM_model_two_lags)
as.numeric(sargan_test_system_model_two_lags$parameter)+
  length(coefficients(system_GMM_model_two_lags)) #75 instruments

#Instrumenting new tests and vaccination to account for the measurement error
system_GMM_model_two_lags_endo<-pgmm(log_new_cases_per_thousand~
                                    lag(log_new_cases_per_thousand,1:2)+
                                    lag(log_new_vaccinations_per_thousand,1:2)+
                                    log_new_tests_per_thousand+avg_temp|
                                    lag(log_new_cases_per_thousand,3)+
                                    lag(log_new_tests_per_thousand,2)+
                                    lag(log_new_vaccinations_per_thousand,4),
                                    data=na.omit(dataset),effect = "twoways",
                                    model=c("twosteps"),transformation = "ld",
                                         index=c("location","date"))
summary(system_GMM_model_two_lags_endo)
#Sargan p-value 0.15 => instruments valid, slight proliferation

mtest(system_GMM_model_two_lags_endo,3)
#p-value 0.44 => No serial correlation of 3rd order => instruments valid

sargan_test_system_model_two_lags_endo<-sargan(system_GMM_model_two_lags_endo)
as.numeric(sargan_test_system_model_two_lags_endo$parameter)+
  length(coefficients(system_GMM_model_two_lags_endo)) #76 instruments
phtest(system_GMM_model_two_lags,system_GMM_model_two_lags_endo)
#Null rejected => Endogeneity is likely present

### Creating a table of results ###
stargazer(diff_GMM_model_two_lags,system_GMM_model_two_lags,
          diff_GMM_model_two_lags_endo,
          system_GMM_model_two_lags_endo,
          diff_GMM_model_one_lag_endo,header = F,title="Estimation_results",
          label="tab:res")
```

Python code

```
import requests
from bs4 import BeautifulSoup
import numpy as np
import pandas as pd

#Obtaining links for all countries
request_countries=requests.get('https://www.weatherbase.
    com/weather/countryall.php3')
request_countries.status_code
soup_countries=BeautifulSoup(request_countries.content)
countries=soup_countries.find_all('a',{'class':'redglow'
    })
countries_href=[i['href'] for i in countries]
countries_href_complete=['https://www.weatherbase.com/'+i
     for i in countries_href]

#Obtaining links for all cities
requests_for_all_countries=[]
for i in countries_href_complete:
    request=requests.get(i)
    if request.status_code==200:
        requests_for_all_countries.append(requests.get(i)
            )
    else:
        print('error'+' '+i)
country_names=[i.text for i in countries]
cities_links_by_country={}
for i in range(len(country_names)):
    soup=BeautifulSoup(requests_for_all_countries[i].
        content)
    cities=soup.find_all('a',{'class':'redglow'})
    cities_links_by_country[country_names[i]]=['https://
        www.weatherbase.com/'+i['href']+'&set=metric' for
        i in cities]

#Obtaining the data for each city in all countries
cities_requests_by_country={}
for i in range(len(country_names)):
    request_list=[]
    for j in cities_links_by_country[country_names[i]]:
        request=requests.get(j)
        if request.status_code==200:
            request_list.append(request)
        else:
```

```
            print('Error'+' '+i+' '+j)
    cities_requests_by_country[country_names[i]]=
        request_list
cities_data_by_country={}
for i in range(len(country_names)):
    data_list=[]
    for j in cities_requests_by_country[country_names[i
        ]]:
        soup=BeautifulSoup(j.content)
        data_all=soup.find_all('td',{'class':'data'})
        avg_temp=[x.text for x in data_all[1:13]]
        final_vector=[]
        for z in avg_temp:
            try:
                final_vector.append(float(z))
            except:
                final_vector.append(z)
        data_list.append(final_vector)
    cities_data_by_country[country_names[i]]=data_list

#Remove empty vectors
cities_data_by_country_final={}
for i in country_names:
    data_list=cities_data_by_country[i]
    cities_data_by_country_final[i]=[x for x in data_list
        if x!=[]]

#Remove too high or too low values
cities_data_by_country_final_finally={}
for i in country_names:
    data_list=cities_data_by_country_final[i]
    new_data_list=[]
    for j in data_list:
        vector=np.array(j)
        vector[vector=='---']=np.nan
        vector=vector.astype(float)
        if sum((vector>56.7)|(vector<-89.2))>0:
            pass
        else:
            new_data_list.append(vector)
    cities_data_by_country_final_finally[i]=np.array(
        new_data_list)

#Computing averages accross cities for each country
monthly_avg_temp={}
for i in country_names:
```

```
        country_data=cities_data_by_country_final_finally[i]
        monthly_avg_temp[i]=np.nanmean(country_data,axis=0)

#Creating a data frame
df=pd.DataFrame(monthly_avg_temp)
df=df.T

#Renaming columns
df.columns=['January','February','March','April','May','
    June','July','August','September','October','November'
    ,'December']
df['Country']=df.index

#Reshaping the data set
df_final=df.melt(id_vars='Country',var_name='Month',
    value_name='Average_Temperature')

#Exporting to csv
df_final.to_csv('avg_temp.csv',index=False)
```