The dynamics of Covid-19: weather, demographics and infection timeline

Renato H. L. Pedrosa

Correspondence:
Dept. of Science and Technology Policy
Institute of Geosciences
University of Campinas – Unicamp
Rua Carlos Gomes, 250
13083-855 Campinas, SP, Brazil
pedrosa@unicamp.br

Abstract

We study the effects of three types of variables on the early pace of spread of Covid-19: weather variables, temperature and absolute humidity; population density; the timeline of Covid-19 infection, as outbreak of disease occurs in different dates for different regions. The regions considered were all 50 U.S. states and 110 countries (those which had enough data available by April 10th. We looked for associations between the above variables and an estimate of the growth rate of cases, the exponential coefficient, computed using data for 10 days starting when state/country reached 100 confirmed cases. The results for U.S. states indicate that one cannot expect that higher temperatures and higher levels of absolute humidity would translate into slower pace of Covid-19 infection rate, at least in the ranges of those variables during the months of February and March of 2020 (-2.4 to 24C and 2.3 to 15g/m^3). In fact, the opposite is true: the higher the temperature and the absolute humidity, the faster the Covid-19 has expanded in the U.S. states, in the early stages of the outbreak. Secondly, using the highest county population density for each state, there is strong positive association between population density and (early) faster spread of Covid-19. Finally, there is strong negative association between the date when a state reached 100 accumulated cases and the speed of Covid-10 outbreak (the later, the lower the estimate of growth rate). When these variables are considered together, only population density and the timeline variable show statistical significance. We also develop the basic models for the collection of countries, without the demographic variable. Despite the evidence, in that case, that warmer and more humid countries have shown lower rates of Covid-19 expansion, the weather variables lose statistical significance when the timeline variable is added.

Introduction

The first reported cases of people showing serious respiratory symptoms, which were later identified as caused by a new variant of *Coronaviridae*, occurred in China during the last month of 2019. Cases soared during the second half of January, going from 80 on January 18th to over 9,000 by the end of the month, with 213 deaths, mostly from acute respiratory insufficiency. At that point, there were 11 cases reported in Japan and only a few other across Asia. By mid-February, there were cases in many other countries, but still in small numbers, and none in Africa or South America, which have most of their territories in the Southern Hemisphere, then under Summer. Australia was one of the few countries in the Southern Hemisphere with cases, but, at that time, all could be traced to travelers arriving from China or other Asian countries. By the end of that month, some countries were already facing an epidemic situation. The first positive case reported by a South American country occurred on Feb. 26th, in Brazil, of a man who had just returned from a trip to Northern Italy. Since then, the virus started to spread, slowly, initially, as in other countries, with all cases being of people arriving from other countries, especially from Italy, or who had direct contact with them. By March 6th, it was recognized that local untraceable transmission was occurring in Brazil and the number of cases increased rapidly. As we will see, the pattern, for the 10 days starting on the day of the 100th case, in Brazil as in many other countries, is exponential to a high degree, with varying growth rate values. During the whole period since the start of the pandemic until the end of March, weather in Brazil has been warm and humid, with temperatures frequently soaring above 32C and absolute humidity never below 10g/m^3. This raises the possibility that warm

and humid weather will not help contain the spread of the virus, as is typical of many viral diseases[0,1,3,4,5]. There are various studies with early, somewhat conflicting, results on those relationships for Covid-19[6,7,8,9,10,11,12,13], as discussed in the National Academies of Science, Engineering and Medicine report on the issue[14].

As indicator of the pace of expansion of Covid-19, we compute as estimate of the pace of the disease's growth the coefficient of the best exponential fit to the evolution curve in the period of 10 days starting when the region reached the 100th case, denoted by *k*, for all 50 U.S. states and 110 countries with enough data (up to April 10th). The control variables are: the average temperature and absolute humidity values during the 25 days starting 15 days before the region reached the 100th case; a couple of timeline variables, the date when the 100th case occurred and the number of days from 1st to 100th cases (for countries and U.S. states); and the (*log* of) population density for the densest county in the state (for U.S. states). We find that the weather variables, albeit significant in single-variable models (but with opposite effects for the two groups of regions, U.S. states and countries), lose significance when the timeline and/or demographic variables are introduced. In the first case, for both groups of regions, the later the date of the 100th case, the slower the initial pace of expansion of Covid-19. In the second one, the higher the population density indicator for the U.S. states, the faster the disease has spread in its initial phase. The model with both variables, only available in the case of U.S. states, furnishes the best estimates, explaining more than 50% of the variability of the growth rate *k*. Finally, the population density also impacts the start of the local transmission phase and how long it took to go from 1st to 100th cases, for U.S. states. The author acknowledges conversations with Aluísio Pinheiro, from the Department of Statistics, Unicamp, about statistical aspects of models employed.

Data and methods

Confirmed Covid-19 cases

Data on Covid-19 cases: databases of reported cases of Johns Hopkins University's Center for Systems Sciences and Engineering (CSSE/JH)[15] for US states the European Centre for Disease Prevention and Control (ECDC/EU)[16] for countries.

Weather data and averages of temperature and absolute humidity

Weather data: NOAA Integrated Database (ISD)[17] of meteorological observations, through the R package "worldmet"[18]. For countries, we used the station nearest to the capital with 100% coverage, when available. In the case of the United States, since most initial cases were reported in the Seattle, in the counties around San Francisco and in New York, we used the averages from those cities. For U.S. states we used the data from station of main airport in the largest city in the state, the more likely place to have had the earliest outbreak. For Brazil, we took the averages for the locations given by the largest airports of São Paulo and Rio de Janeiro, which were the largest cities in states with at least 20 cases on the day of the 100th case.

From NOAA weather data, the absolute humidity Ah (g/m^3) is computed from temperature T (C) and relative humidity Rh (%) using the following approximation formula derived from the Clausius-Clapeyron equation[8,19,20]:

$$Ah = \frac{13.247\,Rh}{273.15 + T}\; exp\left(\frac{17.67\,T}{T\,+\,243.5}\right) \quad (1)$$

The averages for both variables were computed during the period, for each country/U.S. state, of 25 days starting 15 days before the day the region reached 100 accumulated cases of Covid-19. This would cover the period of incubation of two weeks between transmission and evidence of the disease and that of 10 days used to estimate the exponential coefficient for the evolution of cases (see Eq. 2 below). Eq. 1, for constant relative humidity, approximates closely an exponential of T. We have run the regression models with both Ah and $log(Ah)$ as the control variable for absolute humidity (see Figs. 1.d/e and 2.d/e for the distributions). The *log*-transformed version usually provided the stronger regression parameters (p-value and R^2) for Ah.

Countries and U.S. states

We have considered the 110 countries and that had at least 10 days of data starting when they had reached the number of 100 accumulated cases, end-date being at most April 10th. All 50 U.S. states had already reached that stage by that date.

Population density of U.S. states

The population density of the densest county in each U.S. state was used as indicator of population density for the state (U.S. Census 2010[21,22]). Covid-19 has typically started its spread in the city/county of highest population or population density, which almost always coincide (in most cases the city is part of a county, in some, the county is part of a larger city, like for New York city). We have tested the models with two variables: the population of the largest county[23] and the density population for the densest one. Both gave essentially the same general results regarding the regression models employed. We chose to use the latter, as it is easier to interpret regarding the transmission of infectious diseases[24]. For the regression models, we have employed the *log* of population density as control variable, as mathematical models of the dependence of the basic reproductive rate (R_0) of infectious diseases and population density[24] show that one must use non-linear scaling when the range of densities is very wide, which is the case here (from 13 to over 25,000 pop/km^2). This causes the original distribution of population density to be very skewed, concentrated on values below 2,000pop/km^2 (Figs. 1.h/i, 3.a/b). As we analyzed various options of transformations, both in terms producing a more evenly spread distribution and for our application, the study of the early rate of growth of cases of Covid-19, it turned out that the *log*-transformation of population density showed not only good distribution properties (Fig. 3.c), but proved to be well adapted as a control variable for the study of growth rate coefficient. We have not found a discussion about this in the literature, but one may argue that there is increasing dampening of social interaction as population density reaches very large values, and that the dampening follows a logarithmic behavior. As an example, one may think of New York city, where the population density is higher than 25,000 *people/km^2* in some areas. That is certainly caused by people living in tall buildings, since, as the footprint of a building is very small, those living in various floors would be collapsed into a very small area, making the density artificially high. That would increase the interactions as people would likely meet in halls and elevators, but not by a rate indicated by the nominal density.

Estimating the growth rate of number of cases

For each country and U.S. state included in the study, we considered the period of 10 days starting on the day the number of cases reached 100 and computed a simple linear regression for $log(N_i)$ as function of day (t), where N_i is the number of accumulated observed cases for region i at a given day of the period, given by the model

$$log(N_i) = C_i + k_i * t + \varepsilon_i. \quad (2)$$

The growth rate indicator will be the estimated k_i, the exponential coefficient, for region *i*. We have explored also 12-day and 15-day periods starting on the 100th-case day, but there was no relevant impact on the results of the models. We preferred the 10-day model as it provides the best fitting parameters for model in Eq. 2. We could also have estimated the growth rate using endpoints, but we wanted to be able to check the fitting of evolution curves to an exponential. Also, when there are jumps in the reported number of cases, which happened often in the early phase of the disease, the endpoints estimate tends to also jump, while the regressed estimate follows a smoother path (see Fig. 6.b for the case of New York State). Thus, even if one of the endpoints in our 10-day window had shown a sudden jump in the number of cases, that would be smoothed out by our methodology. The choice of 100 accumulated cases to start the analysis is related to when it is expected that local transmission would be under way, as that is when an exponential increase is expected to start. The point is that the first 100 cases for each country or state typically happen within one or a few communities, so that one expects that local transmission is already at work. It has been estimated that when a U.S. county has reached 20 cases, the chances that there is ongoing local transmission are 99%[25,26]. For example, for the United States, the date of the 100th case was March 3rd. On that date, Washington and California led the country, with 27 and 25 cases, respectively. Within the 10-day period used in the estimation, New York had also become a hotspot for cases, and on the last day of the 10-day period, it was second, and other states had surpassed 20 cases. Similarly, for Brazil, the cities of São Paulo and Rio de Janeiro had already had at least 20 confirmed cases when Brazil reported its 100th case.

There are two hypotheses for the estimate k_i to be representative of the speed of transmission of the novel coronavirus in a community. First, *that the way the cases were accounted for did not change significatively along the 10-day period used for the models*. The evidence from the estimates corroborates that assumption, as the values of R^2 for regressions (Eq. 2) are mostly above 0.95 (Figs. 1.b, 2.b). We discuss the sensitivity of our models w.r.t. the R^2 values of regressions given by Eq. 2, in the section on models. The second one is that *the number of positively tested cases of people with various levels of symptoms, or of hospitalized cases, are constant fractions of the total number of people infected (of which, many, are asymptomatic)*. Some countries have tested all people with any symptom, others only those hospitalized (the case for Brazil), and others with in-between levels of symptoms. The model (Eq. 2) employed to estimate the coefficients k_i, under those two hypotheses, would not depend on the alternatives considered by different countries (or states) regarding how they measured confirmed cases.

Modelling the early dynamics of Covid-19

Using the data described above, we consider the following linear regression model for the U.S. states:

$$k = C_0 + C_1 * AvTemp + C_2 * log(AvAh) + C_3 * Day_{100} + C_4 * Days_{1to100} + C_5 * log(PopDens) + \varepsilon, \quad (3)$$

where *k* is the exponential growth rate coefficient estimate (Eq. 2), *AvTemp* is the average temperature, *AvAh* is the average absolute humidity, Day_{100} is the date when region reached 100 cases, $Days_{1to100}$ is the number of days from 1st to 100th case and *PopDens* is the population density of county with highest such value in each state. For the collection of 110 countries we used the same model without the last regressor. We also tested the model for *Ah*, but *log(Ah)* showed consistently better regressing properties than the linear case.

Software: *R*, *ggplot2* and *worldmet* packages.

Preliminary analysis

Figs. 1/2 present the boxplots for the distributions of all relevant variables, for U.S. states and countries, respectively. In Fig. 1.d/e, 2.d/e, one can see how the *log*-transformation of the values for absolute humidity improve the distributions' characteristics. Figs. 1.h.i, 3 show how the log-transform of population density improved the distribution, dramatically in this case. In the case of the variable giving the number of days between the dates of 1st and 100th cases, there are five states (Arizona, California, Illinois, Massachusetts and Washington) which showed outlier character w.r.t. that variable. They took at least 40 days to evolve from 1 to 100 cases, while all other 45 states took at most 23 days. Removing those 5 states from the analysis not only improved the variable distribution, but it resulted as a significative individual regressor of the dependent variable (k, Eq. 2), as we will see in the modeling section. From Figs. 1.a and 2.a, we see that k is spread along a very wide set of values, from almost 0 to almost 0.4. For countries, mean = 0.169, median = 0.155 and SD = 0.082. For U.S. states, mean = 0.208, median = 0.215 and SD = 0.068. The values of R^2 for regressions using Eq. 2 were typically above 0.90 and most above 0.95 (Figs. 1.b, 2.b).

To visualize how well the exponential model for the growth rate k fits the actual trajectories, Fig. 4 displays the exponential prediction and the curves for actual accumulated number of positive Covid-19 cases, for selected U.S. states. For countries, the behavior is similar, we omit the graphs. Figs. 5.a-d show the daily behavior of the two weather variables along the 25-day periods for a choice of U.S. states and countries. Both variables show wider range of values for the U.S. states, compared to the countries displayed. In some cases, the range of temperatures was more than 15C, and over 10 g/m^3 for absolute humidity. For countries the ranges for individual countries are typically smaller, with a few exceptions. The ranges tend to be narrower for regions with higher values, for both variables. It is important to observe that there is no trend regarding when the averages were computed and the actual values for weather variables. Fig. 7 shows the evolution curves of cases for U.S. states for the 10-day period used to estimate the growth rate k, by temperature. It shows that there is a wide range of values of k, and no obvious trend in the temperature averages w.r.t. k. The graph for countries, not included, is similar.

It is possible to relate the values of k with the *basic reproduction number*, R_0, the estimate of the average number of new infections generated by an infectious person, which has been subject of much investigation since the start of pandemic[27]. It has been estimated that R_0 for the early outbreak in China was between 4.7 to 6.6, derived from a growth rate of 0.29[28]. Our estimate for the growth rate k for China is 0.334, but if we start on Jan. 13th (instead of Jan. 19th), and compute k using the next 12 days of evolution, we get k=0.291. If the value of k is reduced from 0.29 down to 0.14, their[28] growth estimate for early February (our estimated k for Feb. 1-10 is same, 0.137), R_0 would drop by 50-59%, to a range between 2.3 to 3.0, certainly due to the containment measures put in place in China in late January. The estimates of k for the U.S. (10-day window Mar. 3-12) is k=0.292 and, for New York State (10-day window Mar. 8-17), is k=0.291, thus both coincide with their[28] estimate for China, and all the above results for R_0 apply equally. In the case of New York, Fig. 6.b indicates that it reached k=0.14 using the 10-day window starting on Mar. 24th, so that, on the 10 days between that date until Apr. 2nd, R_0 had an average value between 2.3 and 3.0, as for China in early February. New York reached the peak for k on the 10-day window Mar. 14-23, at 0.441 (Fig. 6.b), implying that the evolution of k during the second half of March was already showing the impact of social distancing measures (schools were closed on Mar. 16th, but other social distancing measures had been adopted earlier), by reducing the number of new infections caused by each infectious person from around 5.6 to about half of that, in a 16-

day period (from 10-day window Mar. 8-17 to that of Mar 24-Apr. 2). Fig. 6.a shows the evolution of k for a few U.S. states around the date when they reached the 100th case. Most states were already showing a decreasing value of k on the period around the date they reached the 100th case, especially those for which Day_{100} came later in March, which reinforces the idea that social distancing practices were already in place in the second half of March, in most states. This will be discussed further in the next section, as we study the dependence of k on timeline variables.

Models and effect of variables

U.S. states

Table 1 presents the results of regressions for the exponential growth constant (k in Eq. 2) using the model given by Eq. 3 and its sub-models: *AvTemp* is the average temperature, *AvAh* is the average absolute humidity, Day_{100} is the date when region reached 100 cases, $Days_{1to100}$ is the number of days from 1st to 100th case and *PopDens* is the population density of densest county for each state. We also include regressions for the timeline variables w.r.t. population density. All parameters are computed for 95% CI. We omit the values of intercepts, as they are not relevant for the analysis. We include F-statistics p-values for the multivariable cases and Shapiro-Wilk test results for the residuals of all regressions with significative coefficients. In the case of multivariable regressions, R^2 is the adjusted value. There are two sets of models, the first one with all states and a second one with the states which are outliers for at least one of the variables removed (list is below Table 1). Figures in Table 1 indicate that the qualitative results are basically the same, even though the second set of models provide stronger regression parameters for all cases. The last model (9#), which gives the association between the lag between $Days_{1to100}$ and population density, has only the version for the restricted set of states, as the one for all states showed very poor parameters.

From the complete models (1/1#), it is observed that the date when state reached 100th case and the log of population density are the only significant variables. One observes that in those two models, the coefficients of the weather variables reverse signs, not unexpected since their confidence intervals include the zero value. Removing one or two of the other three variables did not change that, only Day_{100} and *log(PopDens)* stayed significant. Using those variables, model 2 explain 53% (68% in the restricted model) of the variability of k, and the residuals satisfy normality to a good level. The coefficient for Day_{100} indicates that if a state reached 100 cases 10 days after another one, k would be reduced by about 0.053 point (0.074 in the restricted model). For example, if k=0.25 for the first, one would expect about k=0.197 (k=0.176) for the second one. That would translate into increasing the time to double the number of cases from 2.8 days to 3.5 days (3.9 days). As an example, New York, which reached 100 cases on March 8th (actually, 106 cases), had 140,000 confirmed cases 30 days later (April 7th), so that the (average) k during the period was about 0.25. As the estimated k for New York for the first 10 days starting on March 8th was 0.291, it means that there was attenuation of the pace of spread of Covid-19 since at least March 17th, which is confirmed by its evolution curve and varying k (Fig. 6.b).

The population density coefficient of model 2 results in that doubling the population density implies that the value of k increases by 0.0162**log*(2) = 0.0112. For example, the states of Iowa and Missouri, which reached the 100th case on Mar. 23rd, had k's estimated as 0.195 and 0.235, respectively, thus a difference of 0.040. Their population densities are 290 and 1,991 people/k^2, respectively, which imply an estimated increase of 0.031 point to the value of k, or about 75% of the actual estimate. This is within the expected confidence intervals of the models, as they predict at most 68% of the variability of k (adjusted R^2, model 2#).

Table 1 - Regression results: U.S. states

Model	Dependent variable	Independent variable(s)	Coefficient	SD	p-value (k)	R^2 (1)	F-stat p-value	Shapiro-Wilk p-value (residuals)
1	Exponential growth coefficient *k*	AvTemp	0.0023	0.0030	0.4455	0.5697	<0.001***	0.5974
		log(AvAh)	-0.0105	0.0495	0.8128			
		Day_100	-0.0055	0.0020	0.0073**			
		Days_1to100	-0.0008	0.0006	0.2200			
		log(PopDens)	0.0147	0.0069	0.0380*			
1#	*k*	AvTemp	-0.0002	0.0030	0.9497	0.6667	<0.001***	0.8576
		log(AvAh)	0.0071	0.0409	0.8637			
		Day_100	-0.0074	0.0028	0.0111*			
		Days_1to100	-0.0020	0.0019	0.2843			
		log(PopDens)	0.0186	0.0073	0.0150*			
2	*k*	Day_100	-0.0053	0.0019	0.0064**	0.5305	<0.001***	0.9233
		log(PopDens)	0.0162	0.0068	0.0214*			
2#	*k*	Day_100	-0.0091	0.0021	<0.001***	0.6803	<0.001***	0.5795
		log(PopDens)	0.0167	0.0067	0.0168*			
3	*k*	AvTemp	0.0036	0.0015	0.0237*	0.1021		0.5967
3#	*k*	AvTemp	0.0040	0.0016	0.0188*	0.1246		0.4608
4	*k*	log(AvAh)	0.0485	0.0226	0.0372*	0.0874		0.7117
4#	*k*	log(AvAh)	0.0522	0.0234	0.0312*	0.1058		0.4471
5	*k*	Day_100	-0.0086	0.0013	<0.001***	0.4738		0.9718
5#	*k*	Day_100	-0.0129	0.0015	<0.001***	0.6490		0.6222
6	*k*	Days_1to100	-0.0002	0.0009	0.8150	0.0012		
6#	*k*	Days_1to100	-0.0081	0.0022	<0.001***	0.2412		0.3311
7	*k*	log(PopDens)	0.0305	0.0049	<0.001***	0.4491		0.6426
7#	*k*	log(PopDens)	0.0385	0.0054	<0.001***	0.5515		0.4897
8	Day_100	log(PopDens)	-2.6895	0.3523	<0.001***	0.5526		**0.0024**
8#	Day_100	log(PopDens)	-2.3979	0.3329	<0.001***	0.5526		0.3039
9#	Days_1to100	log(PopDens)	-0.8578	0.4644	0.0718	0.0751		0.6

(1) Adjusted R^2 for multivariable regressions

Model omits Arizona, California, Illinois, Massachusetts, New York and Washington (see text)

Table 2 - Effect of *k* on the evolution of COVID-19 cases

k	0.05	0.10	0.15	0.20	0.25	0.30	0.35
days to double	13.9	6.9	4.6	3.5	2.8	2.3	2.0
Cases 1 month after reaching 100 cases	448	2,009	9,002	40,342	180,804	810,308	3,631,550

No other 3- or 2-variable models with *k* as dependent variable provided new relevant results. The 1-variable models (3/3# to 7/7#) show significative results, except for model 6 (case with relevant outliers, as discussed). The possible surprise is that both average temperature and average absolute humidity have positive coefficients, individually, which is confirmed by the scatter plots in Fig. 8. Those behaviors are actually a consequence of the fact that many colder states, especially the less densely populated ones, started their outbreaks later, thus affecting these results, according to models 2/2#. Figs. 8.a-e include scatter plots and trend lines for *k* and the control variables, to illustrate the above results, with trend lines and parameters from the models for all states.

Models 8/8# show that (log of) population density has significant impact on when a state reached 100 cases. This is expected, as not only lower population density slows the pace of Covid-19 infection, as we have seen, but one may also think that the introduction of the virus would have been delayed in such states (which, in fact, is the case, checking the date of first reported case for each state), and thus it would be the case that social

distancing was already occurring when that happened. The estimate given by the coefficient in model 8 is that doubling the population density would make the day of 100th happen about 1.9 day earlier. Fig. 9.a shows how strong the association is, with the exception of the state of Washington, which was the first one to have an outbreak and thus does not follow the general trend, as its densest county, which had the very first cases in the U.S., has relatively low population density (King County, 352 pop/km^2). Relaxing statistical requirements, model 9# shows that the time lag between first and 100th cases is reduced by about 0.6 day when the population density doubles. Fig. 9.b shows that the association in this case is in fact weak, not just caused by outliers. Still, there is a clear trend of shortening the period from case 1 to case 100 as population density increases (New York is not an outlier in this case, as its value was 5 days, but we kept the same set of states in the restricted set for consistency).

Countries

Table 3 shows regression results for the group of 110 countries in our database, starting with complete model 1, which is same as given by Eq. 3, dropping the demographic variable. Analogously to the U.S. states' case, the variable Day_{100} shows statistical significance in the complete model. We have two versions, model 1 one with China and model 2 without. The reason is that China is a clear outlier for that variable (see Fig. 2.f), having reached the 100th case on Jan. 19th, more than a month before the second country to do so (South Korea, on Feb. 20th). In any case, the results are essentially the same. In the complete models (1,2), if we relax the statistical requirements for significance, temperature would contribute positively and absolute humidity, negatively, to the estimate of *k*. The same is true for model 3, for these two variables only. This behavior is similar to that of the weather variables in the case of U.S. states (Table 1, models 1/1#, and model with *AvTemp* and *log(AvAh)* as control variables, not included in Table 1).

Table 3 - Regression results: all countries

Model	Dependent variable	Independent variable(s)	Coefficient	SD	p-value	R^2 (1)	F-stat p-value	Shapiro-Wilk p-value (residuals)
1	Exponential growth coefficient *k*	AvTemp	0.0013	0.0020	0.5010	0.3293	<0.001***	0.9209
		log(AvAh)	-0.0378	0.0323	0.2440			
		Day_100	-0.0042	0.0007	<0.001***			
		Days_1to100	-0.0004	0.0005	0.4500			
1#	*k*	AvTemp	0.0019	0.0020	0.3320	0.3232	<0.001***	0.6978
		log(AvAh)	-0.0493	0.0325	0.1330			
		Day_100	-0.0049	0.0008	<0.001***			
		Days_1to100	-0.0003	0.0005	0.5030			
2	*k*	AvTemp	0.0004	0.0023	0.8670	0.1011	0.0033**	**0.0013**
		log(AvAh)	-0.0511	0.0373	0.1740			
3	*k*	AvTemp	-0.0026	0.0008	0.0020**	0.0853		**0.0045**
4	*k*	AvAh	-0.0042	0.0013	0.0014**	0.0910		**0.0010**
5	*k*	log(AvAh)	-0.0452	0.0130	<0.001***	0.1017		**0.0015**
6	*k*	Day_100	-0.0045	0.0006	<0.001***	0.3280		0.8427
6#	*k*	Day_100	-0.0051	0.0007	<0.001***	0.3166		0.3170
7	*k*	Days_1to100	<0.0001	0.0006	0.998	<0.0001		

(1) Adjusted R^2 for multivariable regressions
Omits China

The 1-variable models show what some studies have reported, that temperature and absolute humidity go along with faster pace of spread of Covid-19, but one must note that the residuals, in both cases, show low *p-values* in the Shapiro-Wilk test for the residuals (see Fig. 10). Until we have more countries satisfying the criterion to be included in the study, we cannot say, for the group of countries with data available for our

models to work, much about weather variables' impact on the early pace of Covid-19 infection. One can say, though, that, similarly to the case for the U.S. states, the later a country reaches the 100th case, the lower the expected pace of spread of the disease is. So far, that variable explains about 30% of the variability of the growth coefficient *k* (models 6/6#). The value for the coefficient in this case is similar to that of the case of U.S. states, the value of *k* is reduced by about 0.05 by delaying the date of when country reaches the 100th by 10 days. The variable $Days_{1to100}$ did not show relevant association with the value of *k* for countries, even relaxing statistical requirements. Fig. 11 presents the scatter plots for *k* and relevant control variables.

Sensitivity of models with respect to R^2 of growth rate estimates

All the above models were run for subgroups of countries/U.S. states with values of R^2 (model of Eq. 2) above the levels of 0.90, 0.95 and 0.97. There were no relevant qualitative differences between those models and the ones employing all countries/U.S. states, just small coefficient and parameter variations.

Summary of results and discussion

U.S. states

- Results for the 50 U.S. states indicate that weather variables (average temperature and absolute humidity for the period of 25 days starting 15 days before the date of 100th case), once one takes into account timeline of the diseases evolution and demographic information, have little effect on the rate of expansion of Covid-19, at least in the early phase of the outbreak, considered in this study as the 10 days starting when state reached 100 cases.
- Individually, both variables show positive association with pace of spread of Covid-19.
- The timeline (day when state reached 100th case) and demographic (population density of county with highest value) variables explain in good measure (over 50%) of the variability of the of the estimate of rate of expansion (*k* in Eq. 2) among U.S. states.
- For each 10 days of delay in reaching the 100th case, the coefficient *k* is reduced by about 0.053 point (0.074 point in the restricted model). Table 2 indicates how relevant it is to try to reduce the value of *k* to keep number of cases from exploding. Fig. 6.b for New York State illustrates that in a real case.
- Eq. 5 shows the effect of the population density of the county of highest value, for a state: doubling the population density variable would imply an expected increase of 0.011 point on the value of *k*.
- Regressions 8/8# in Table 1 indicate that the population density also impacts significantly when the 100th case occurred, by making it to occur earlier if population density is higher (1.9 day by doubling the population density). And, relaxing statistical significance requirements, model 9# indicates that doubling the population density reduces the time between the 1st and 100th cases by 0.6 day.

Countries

- For the 110 countries in our database, the date when the 100th case occurred is the only significant variable for the complete model (Table 3, models 1/1#).
- Using the individual regression for that variable, for each 10 days of delay in that occurrence, one reduces the value of *k* by 0.045 (0.051 for model 1#), which is about the same as in the case of the U.S. states. About 33% of the variability of *k* is explained by the date the country reached the 100th case.
- Individually, the weather variables show significative negative association with the pace of Covid-19 expansion, but the models suffer from statistical limitation, as the normality of residuals is not guaranteed (Table 3, models 3-5).

The results summarized above indicate that the weather variables considered in this study do not seem to be relevant determinants for the pace of early spread of Covid-19. Even when temperature and absolute humidity are considered in isolation, the results for U.S. states and for the group of countries in our database show opposite behavior. The population density of densest county U.S. states imply faster paces of expansion of the disease. This result is expected, as higher population density implies various characteristics of communities that would help a high level of contact among people. Another aspect is that the denser counties include or are part of cities with higher levels of circulation of outside travelers. But even taking those aspects into account, the best model(s) for U.S. states includes the effect of the date when the 100th case was reached, which helps explain the variability of *k*. For countries, that is also the most important factor to explain the growth rate of the disease. This is likely a result of measures that people and governments started to take as the seriousness of the disease became more evident. An alternative explanation would be that the virus is losing strength, but there is no evidence of that, at least so far.

Table 4 - Countries/U.S. States with AvTemp > 15C and k > 0.180

Country/State	Region	AvTemp	AvAH	k	R2	Day1	Day100
Florida	USA	23.9	13.4	0.319	0.972	2020-03-02	2020-03-15
Brazil	SoAmerica	23.2	16.6	0.307	0.993	2020-02-26	2020-03-15
Louisiana	USA	21.7	14.1	0.295	0.982	2020-03-11	2020-03-16
Ecuador	SoAmerica	19.8	10.1	0.291	0.974	2020-03-01	2020-03-18
Texas	USA	16.9	11.3	0.269	0.962	2020-03-05	2020-03-17
Nigeria	SSahAfrica	29.1	20.2	0.261	0.981	2020-02-28	2020-03-31
South Africa	SSahAfrica	18.3	10.1	0.259	0.993	2020-03-06	2020-03-19
Arizona	USA	18.0	7.0	0.258	0.976	2020-01-26	2020-03-21
Georgia	USA	15.2	9.6	0.253	0.976	2020-03-03	2020-03-16
Thailand	Asia/Pacific	29.4	19.1	0.248	0.942	2020-01-13	2020-03-16
Alabama	USA	17.7	11.0	0.243	0.985	2020-03-13	2020-03-20
Israel	MEast/NAfrica	16.2	9.3	0.238	0.975	2020-02-22	2020-03-14
Malaysia	Asia/Pacific	29.1	21.1	0.233	0.952	2020-01-25	2020-03-10
Domenican Rep.	CentAm/Carib	23.7	17.9	0.220	0.887	2020-03-02	2020-03-22
Chile	SoAmerica	21.1	8.8	0.220	0.992	2020-03-04	2020-03-17
Saudi Arabia	MEast/NAfrica	21.5	5.4	0.220	0.977	2020-03-03	2020-03-16
Panama	CentrAm/Carib	29.9	18.5	0.219	0.993	2020-03-10	2020-03-18
Indonesia	Asia/Pacific	27.7	21.1	0.211	0.978	2020-03-02	2020-03-16
New Zeland	Asia/Pacific	17.8	11.2	0.210	0.979	2020-02-28	2020-03-23
Pakistan	Asia/Pacific	16.6	9.3	0.208	0.941	2020-02-27	2020-03-17
Argentina	SoAmerica	22.6	14.8	0.201	0.995	2020-03-04	2020-03-20
Australia	Asia/Pacific	21.7	13.6	0.200	0.994	2020-01-25	2020-03-10
India	Asia/Pacific	21.3	11.8	0.200	0.983	2020-01-30	2020-03-17
South Carolina	USA	17.7	11.4	0.194	0.976	2020-03-07	2020-03-20
Mexico	NoAmerica	19.6	5.9	0.188	0.984	2020-02-29	2020-03-19

Some final comments on the question if warmer and more humid weather would help reduce the dissemination of Covid-19. Besides the results of models in this study, it is relevant to look at cases of countries/U.S. states for which one has both higher temperatures and absolute humidity values and also higher levels for the coefficient *k.* Table 4 presents data for countries/U.S. states with average temperatures above 15C and *k* above 0.18, which is the average of *k* for all countries and U.S. states (removing the U.S. from the countries' database). It also includes the R^2 estimates for the determination of *k* and the dates when 1st and 100th cases were reported. All regions in the table had the date of 100th case on Mar. 10th or later, so that they are not a group of countries with very early Day_{100} variable, which could have impacted positively its value of *k*, as predicted from our models. Regions with highest values of *k*, above 0.25, include, in descending order of *k*: Florida, Brazil, Louisiana, Ecuador, Texas, Nigeria, South Africa, Arizona and Georgia (state). Of those, Florida, Brazil, Louisiana and Nigeria had average temperatures above 20C, and average humidity values above 11g/m^3, with Nigeria's and Brazil's above 15g/m^3. Except for South Carolina and Mexico, all regions

included in the table show values of *k* between 0.20 and 0.25, which implies a rate of doubling cases of at most 3.5 days (Table 2). Countries in that group with temperatures above 25C include Thailand, Malaysia, Panama and Indonesia, all with high levels of absolute humidity (>18g/m^3). Preliminary data for Brazilian states show that, like their U.S. counterparts, warmer and more humid weather seem to imply faster pace of spread of Covid-19. For Amazonas (in the heart of the Amazon rainforest), Ceará and Pernambuco, states with fast early growth of Covid-19 cases, average temperatures were above 27C and absolute humidity levels above 20g/m^3, during the month of March.

There are also cases of countries in the database with high average temperatures and levels of absolute humidity and low values for *k*, but we think there are higher chances that there are possible reasons for that: one possibility is the lack of reporting the development of Covid-19 by local authorities; another, that some countries cannot apply enough testing and, therefore cannot maintain a reliable sequence of reports, possibly impacting the rate of expansion negatively. Anyway, the existence of many countries and U.S. states (and Brazilian states) with warm and humid weather and fast expansion rates of Covid-19 indicates that the way for countries and regions to keep the evolution of expansion of the disease under control, within the capabilities of their health systems (at least until the development of vaccines or effective therapeutics is successful), is to keep employing social distancing policies, which is predicted by modelling[29] and seems to be working effectively in all countries and states that have adopted them.

Limitations: the above results are preliminary in scope and depend on the quality of available data. As more countries and change in seasons bring new information for analysis, models may be updated and further developed.

References

1. Cai, Q.-C., J. Lu, Q.-F. Xu, Q. Guo, D.Z. Xu, Q.-W. Sun, H. Yang, G.-M. Zhao, Q.W. Jiang (2007). Influence of meteorological factors and air pollution on the outbreak of severe acute respiratory syndrome. Public Health 2007; 121: 258–65.
2. Price, R.H.M., C. Graham, S. Ramalingam (2019). Association between viral seasonality and meteorological factors. Sci Rep 2019; 9: 1–11.
3. Altamimi A., A.E. Ahmed. Climate factors and incidence of Middle East respiratory syndrome coronavirus (2019). J Infect Public Health 2019.
4. Sun Z., K. Thilakavathy, S.S. Kumar, G. He, S.V. Liu (2020). Potential Factors Influencing Repeated SARS Outbreaks in China. Int J Environ Res Public Health 2020; 17: 1633.
5. Cohen, J. (2020). Dozens of diseases wax and wane with the seasons. Will Covid-19? Science 2020; 367: 1294–7.
6. Luo, W., M.S. Majumder, D. Liu, C. Poirier, K.D. Mandl, M. Lipsitch, M. Santillana (2020). The role of absolute humidity on the transmission rates of the Covid-19 outbreak. https://doi.org/10.1101/2020.02.12.20022467.
7. Sajadi, M. M., P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, A. Amoroso (2020). Temperature, humidity and latitude analysis to predict potential spread and seasonality for Covid-19. http://dx.doi.org/10.2139/ssrn.3550308
8. Bukhari, Q. and Y. Jameel (2020). Will Coronavirus Pandemic Diminish by Summer? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3556998 (posted 2020/03/19).
9. Ficetola, G. F., and D. Rubolini. 2020. Climate affects global patterns of Covid-19 early outbreak dynamics. https://doi.org/10.1101/2020.03.23.20040501
10. Wang, J., K. Tang, K. Feng, and W. Lv. 2020. High temperature and high humidity reduce the transmission of Covid-19. http://dx.doi.org/10.2139/ssrn.3551767
11. Islam, N., S. Shabnam, A.M. Erzurumluoglu. Temperature, humidity, and wind speed are associated with lower Covid-19 incidence. https://doi.org/10.1101/2020.03.27.20045658
12. Notari, A. Temperature dependence of Covid-19 transmission. https://doi.org/10.1101/2020.03.26.20044529

13. Qi, H., S. Xiao, R. Shi., M.P. Ward, Y. Chen, W. Tu, Q. Su1, W. Wang, X. Wang, Z. Zhang (2020). Covid-19 transmission in Mainland China is associated with temperature and humidity: a time-series analysis. https://doi.org/10.1101/2020.03.30.20044099
14. National Academies of Science, Engineering, Medicine (2020). Rapid Expert Consultation on SARS-CoV-2 Survival in relation to Temperature and Humidity and Potential for Seasonality for Covid-19 Pandemic (April 7, 2020). Washington, DC: The National Academies Press. http://nap.edu/25771 https://doi.org/10.17226/25771
15. Johns Hopkins Center for Systems Sciences and Engineering. Github CSSEGISandData/Coronavirus site. https://github.com/CSSEGISandData/Covid-19
16. European Centre for Disease Prevention and Control, European Union. https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-Covid-19-cases-worldwide (downloaded 2020/04/11).
17. National Oceanic and Atmospheric Administration, Integrated Surface Database. https://www.ncdc.noaa.gov/isd (downloaded 2020/04/12).
18. Carslaw, D. Worldmet R software package, Github. https://github.com/davidcarslaw/worldmet
19. Iribarne, J. V., W. L. Godson (2012). *Atmospheric Thermodynamics*. Springer Science & Business Media.
20. Bolton, D. The Computation of Equivalent Potential Temperature (1980). *Mon. Weather Rev.* **108**, 1046–1053.
21. County population density https://archive.vn/20150408083711/http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_SF1_G001&prodType=table (access 2020/03/25, access closed since 2020/03/31, available in reference 22).
22. Wikipedia, County Statistics, https://en.wikipedia.org/wiki/County_statistics_of_the_United_States (access 2020/04/13).
23. County population, U.S. Census. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html#par_textimage_739801612 (access 2020/04/13).
24. Hu H., K. Nigmatulina, P. Eckhoff. The scaling of contact rates with population density for the infectious disease models. Math. Biosci. 2013; 244(2): 125-34. https://www.researchgate.net/publication/236691140_The_Scaling_of_Contact_Rates_with_Population_Density_for_the_Infectious_Disease_Models
25. Javan, E., Dr. S. J. Fox,Dr. L. A. Meyers (2020). Probability of current Covid-19 outbreaks in all US counties. https://cid.utexas.edu/sites/default/files/cid/files/Covid-risk-maps_counties_4.3.2020.pdf?m=1585958755 (access 2020/04/06).
26. New York Times (2020/04/03), https://www.nytimes.com/interactive/2020/04/03/us/coronavirus-county-epidemics.html
27. Liu, Y., A.A. Gayle, A. Wilder-Smith, J. Rocklöv (2020). The reproductive number of Covid-19 is higher compared to SARS coronavirus. Research Letter, J. Travel Med, 2020, 1-4. doi:10.1093/jtm/taaa021
28. Sanche S., Y.T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, R. Ke (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis, 2020/July. https://doi.org/10.3201/eid2607.200282
29. Bendtsen Cano, O., S. Cano Morales, C. Bendtsen. (2020) Covid-19 Modelling: the Effects of Social Distancing. https://doi.org/10.1101/2020.03.29.20046870

Figures

Figure 1.a-i. - Boxplots for regression variables and for R^2 for the exponential regression for k (Eq. 2). U.S. states, by region. Boxplot for population density (h) does not include value for New York (26,822 persons/km^2), to allow for better visualization of distribution. Data: CESS/JHU, NOAA/USDC, U.S. Census (2010).

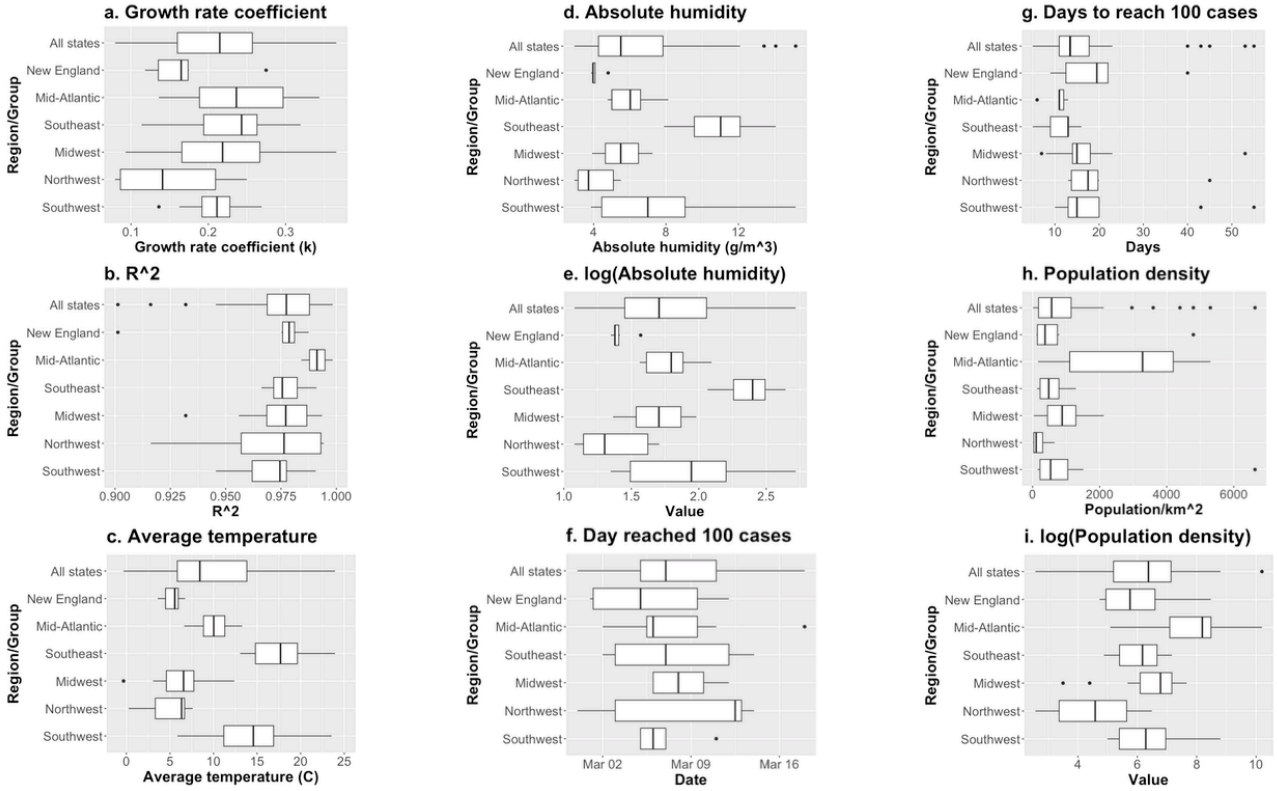

Figure 2.a-g. - Boxplots for regression variables and for R^2 for the exponential regression for k (Eq. 2). Countries, by region. Data: ECDC/EU, NOAA/USDC.

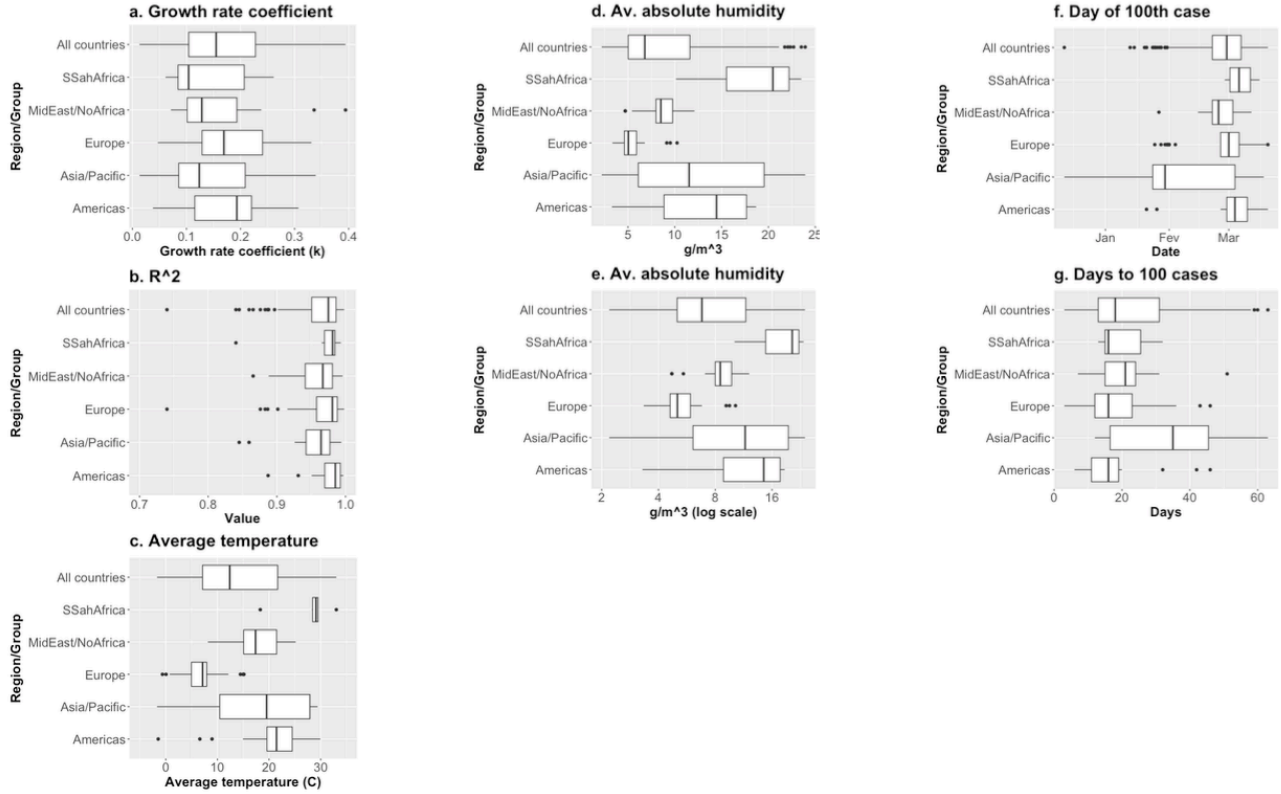

Figure 3: Population density of state's densest county: linear and *log*-transformed. Data: CESS/JHU, NOAA.

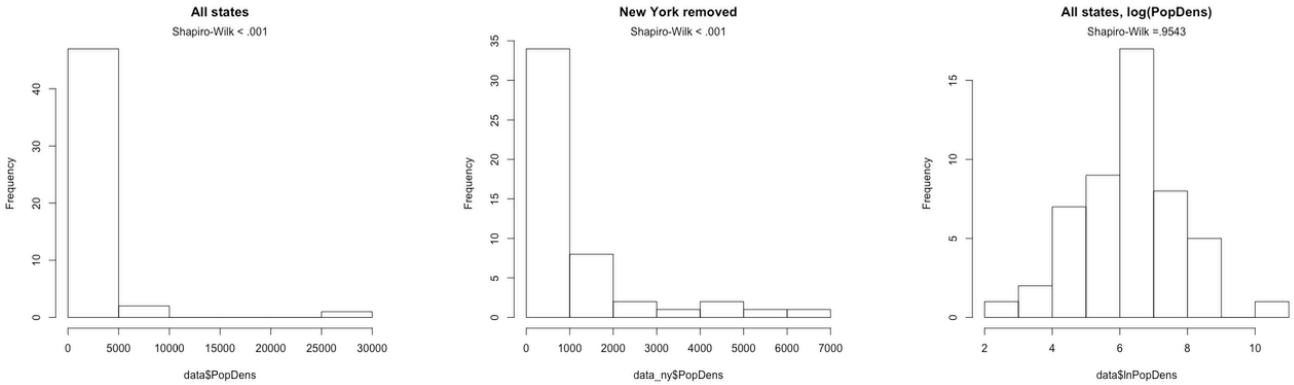

Figure 4: Actual and regressed (Eq. 2) curves of number of confirmed cases, selected U.S. states, 10 days starting on the day of 100th case. Values of k and R^2 for the regressed exponentials. Data: CESS/JHU, NOAA.

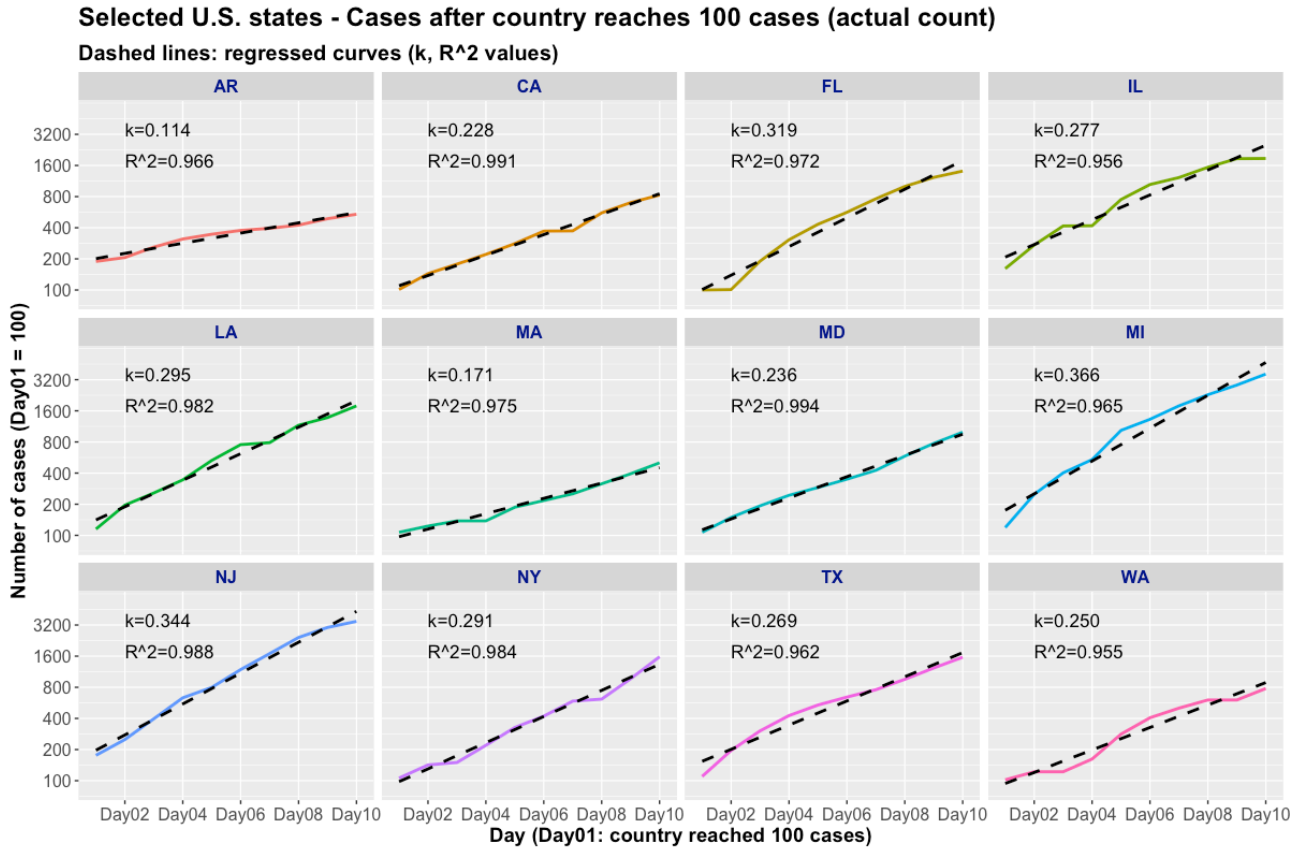

Figure 5.a-d: Daily average temperature and absolute humidity, U.S. states and countries. Period of 25 days, starting 15 days before state/country reached 100th case. Data: NOAA/USDC, CSSEC/Johns Hopkins, ECDC/EU.

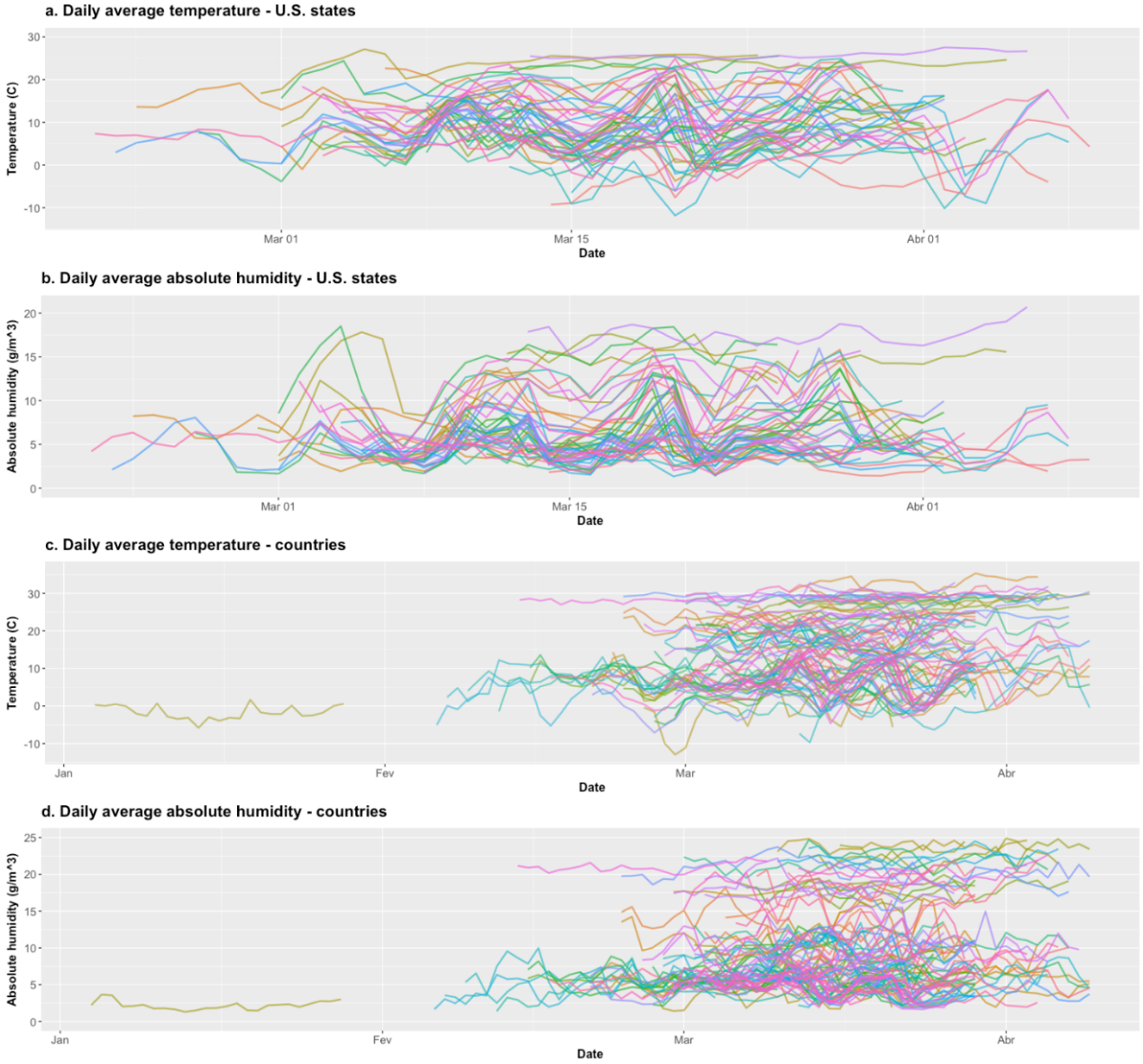

Figure 6.a-b – a. Evolution of *k* from 5 days before to 5 days after Day_{100}, selected states. b. New York State: number of cases, March/05-April/10, *k* and endpoints estimate. Values marked on the first day of 10-day window used for estimates. Data: CSSE/JHU.

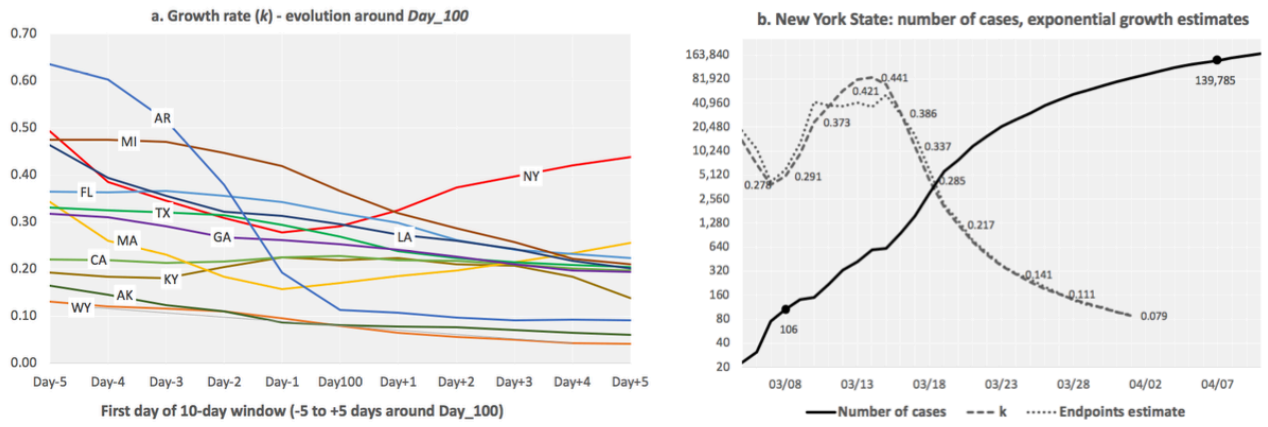

Figure 7 – Number of confirmed cases for 10 days starting when U.S. state reached 100 cases (normalized count, Day01 = 100), by temperature. Dotted lines: exponential curves with number of days to double the number of cases and value of *k*. Data: CSSE/JHU, NOAA/USDC.

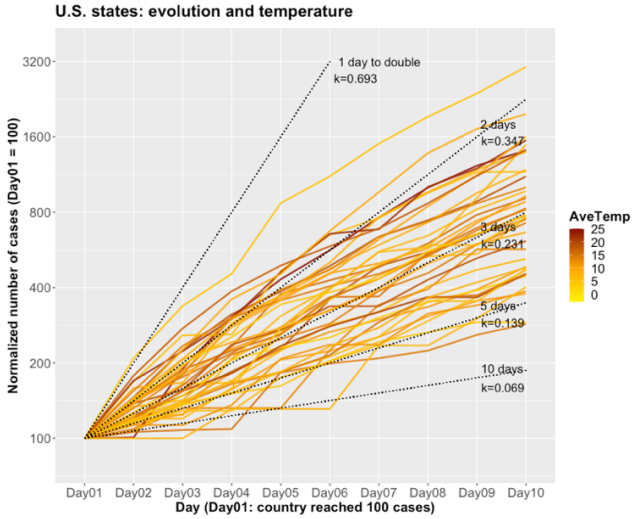

Figure 8.a-e – Scatter plots of exponential growth rate (*k*) and control variables. U.S. states. Data: CSSE/JHU, NOAA/USDC, U.S. Census (2010).

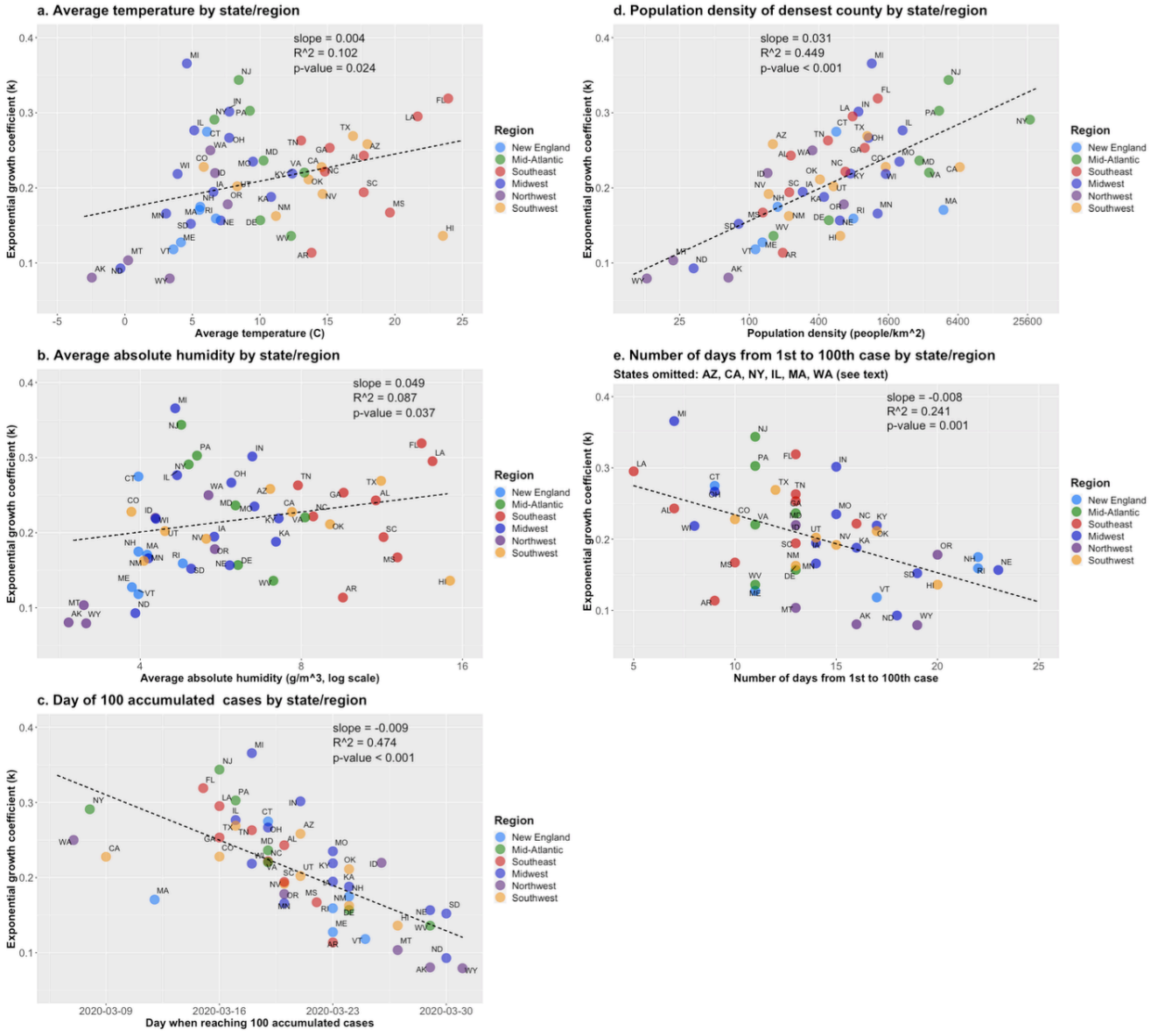

Figure 9.a-b – Scatter plots of timeline variables and population density. U.S. states. Data: CSSE/JHU, U.S. Census (2010).

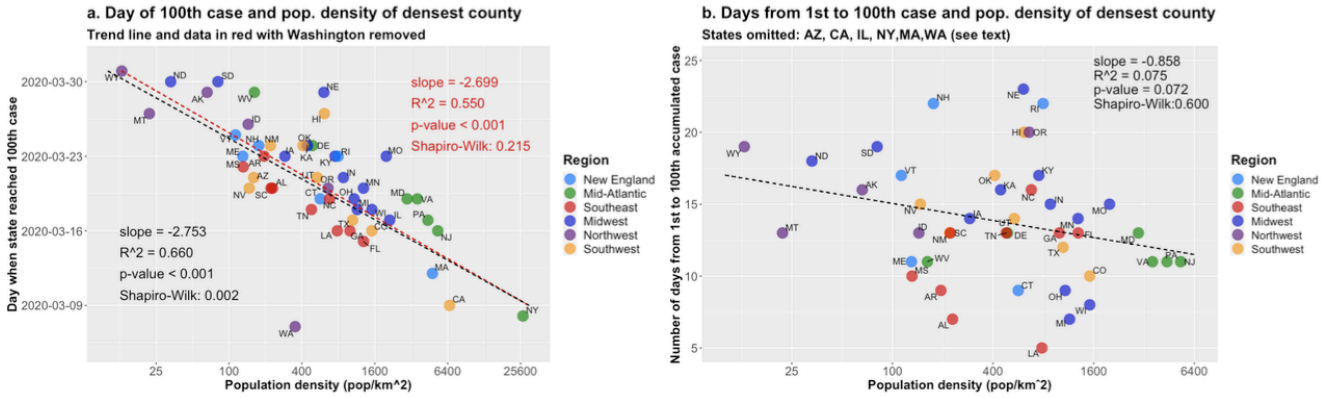

Figure 10 – Residuals of regression Table 3.2.

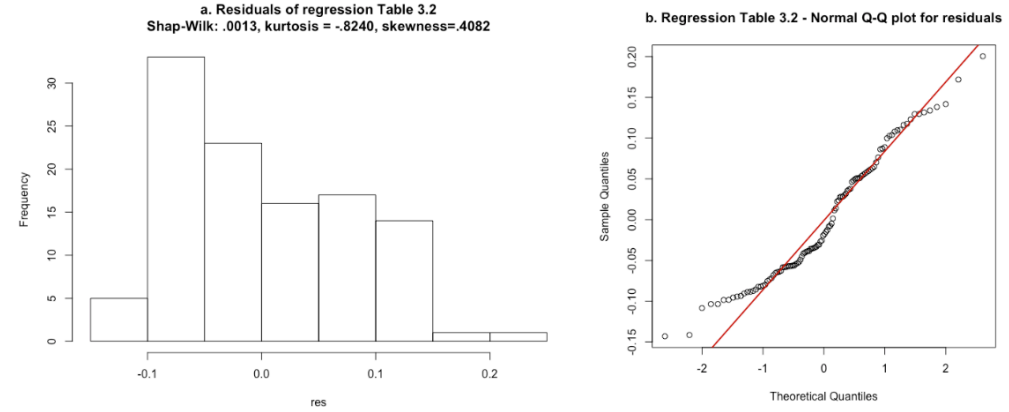

Figure 11.a-c – Scatter plots, exponential growth rate (*k*) and control variables. Countries. Plot c. does not include China (see text). Data: ECDC/EU, NOAA/USDC.