The Effect of Temperature on the Incidence Rate of COVID19

Emanuele Massetti*
School of Public Policy, Georgia Institute of Technology, Atlanta, GA, United States
emanuele.massetti@gatech.edu
https://orcid.org/0000-0003-1659-7051

Eric Dobbie
School of Economics, Georgia Institute of Technology, Atlanta, GA, United States
Edobbie3@gatech.edu
https://orcid.org/0000-0002-5763-6531

Wei Yang
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States
yang.wei@gatech.edu
https://orcid.org/0000-0001-8460-3121

Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States
Kpaynabar3@gatech.edu
https://orcid.org/0000-0002-6906-3611

Abstract

In this paper, we exploit random variation of daily temperature in the United States at both state and county level, from March 1st to October 31st 2020, to study if temperature has a significant effect on COVID19 incidence rate. We find that warmer than average days lead to a lower incidence rate, seven days later. A week in which temperature is consistently one standard deviation above the monthly average in all US states causes 17,754 fewer cases at national level, seven days later. Other weather variables do not have a significant and robust effect on the incidence rate. The effect of temperature is heterogeneous over space and time.
Main

The effect of temperature on the diffusion of COVID19 is still poorly understood. Ambient temperature may affect the incidence rate of COVID19 in two ways. The first is by means of purely environmental effects of temperature on the virus. The second is by changing behavior in ways that lead to different exposure to the virus. Limited and preliminary epidemiological evidence in controlled environments suggests that SARS-CoV-2 survives better at low temperatures and extreme humidity conditions.\(^1\) However, environmental conditions can be very different from those in laboratories and human behavior is important in explaining how temperature affects the diffusion of many infectious diseases.\(^2\) To provide guidance to public health officials, it is necessary to study how temperature affects the diffusion of SARS-CoV-2 and the emergence of COVID19 – the disease caused by the novel coronavirus – in uncontrolled environments with realistic environmental and behavioral conditions. The empirical challenge is to separate the effect of temperature from other weather variables and from factors that change human behavior. Here, we exploit random variation of daily temperature in the United States at both the state and county level, from March 1\(^{st}\) to October 31\(^{st}\) 2020, to study if temperature has a significant effect on COVID19 incidence rate. We find that warmer than average days lead to a lower incidence rate, seven days later. A week in which temperature is consistently one standard deviation above the monthly average in all US states causes 17,754 fewer cases at national level, seven days later. Other weather variables do not have a significant and robust effect on the incidence rate. The effect of temperature is heterogeneous over space and time. During the summer months, higher temperatures lead to a lower incidence rate or have an insignificant effect. The beneficial effect of temperature is larger in areas that are on average colder. Warmer days lead instead to a higher incidence rate in some areas with larger than average temperature. We also find that the larger is
the deviation from average temperature, the larger is the effect on the incidence rate. This suggests a non-linear, U-shaped relationship between temperature and the incidence rate. Using high-frequency and high-resolution mobile phone location data, we find suggestive evidence that the mechanism that drives this result is social behavior. Warmer than average days push people to spend more time outdoors, hence reducing the number of cases. This evidence corroborates the intuition of public health experts that the winter months are particularly challenging for the containment of SARS-CoV-2. Our estimates should be interpreted as a lower bound to the true effect of temperature on the incidence rate of COVID-19 because we use restrictive assumptions to avoid confounding effects. Our study relies on a relatively parsimonious set of data that is publicly available. We expect that our method will be replicated using data from other countries to check if our results hold under alternative environmental, policy and socio-economic conditions.

Our main premise in this study is that the analysis of temperature effects on the incidence rate of COVID-19 should capture both direct (environmental) and indirect effects (behavioral) of temperature. As individual behavior, policies and many evolving conditions cannot be replicated in laboratories, we rely on data-driven methods.

We begin with building a panel of daily observations at both state and county levels in the United States, from March 1st to October 31st 2020. We use publicly available data on the number of positive COVID-19 tests, daily weather variables, and data on movement patterns from GPS position of cellular devices.
Starting from a few clusters, the novel coronavirus has quickly spread to all states and all counties. This diffusion pattern has not been random or consistent over time or space. Geographic and economic proximity to the first clusters, socio-economic characteristics, and policies, all explain the diffusion pattern of the virus. In the ideal study, researchers are able to control all these confounding factors to reveal the sole effect of temperature. Unfortunately, this seems to be an insurmountable challenge at this time. We start from a less ambitious goal and we aim at explaining the role that small but random variations of temperature may have in explaining the incidence rate of COVID19 at the state and county levels.

In our main specification, we remove from all variables their state (or county) by month averages, further subtracting US averages in each day of the week to remove the weekly cycle observed in reports of COVID19 cases (see Methods). As temperature is correlated with other weather variables that may affect the incidence rate directly and indirectly, we control for precipitations, cloud coverage and relative humidity. To account for spatial effects in the diffusion of the virus, we control for the average incidence in neighboring states (or counties). We apply a moving average filter with a three-day window to smooth-out short-term fluctuations. We apply a seven-day lag to all the control variables to mirror the time between contagion and official recording of test results. The seven-day lag is chosen to be at the center of the one-to-fourteen-day interval between contagion and appearance of COVID-19 symptoms identified by the WHO. We verify the choice of the lag and the choice of the control variables using the Akaike Information Criterion (AIC) and variable selection methods.
We limit the range of variation used to identify the effect of temperature on the incidence rate to within state-by-month random temperature oscillations to remove all the possible confounding effects using state-by-month fixed effects. We implicitly control for all state-level differences (e.g. public health, demographics, and politics) and all state-by-month changes of behavior (e.g. summer vacations, school calendar, public health regulations). For counties, we use county fixed effects and state-by-month fixed effects. This method leads to an effective control of the trend in the diffusion of the virus. The downside of our empirical strategy is that we cannot forecast the effect of large temperature changes across seasons. In robustness tests we use state-by-week fixed effects to control for public health policies and other factors that may change within each month.

**Warmer days reduce the incidence rate one week later**

Our results indicate that one additional °C of temperature over three days significantly reduces the daily incidence rate of COVID19 by 0.263 cases per 100,000 people one week later (column (1) in Table 1 in Extended Data - ED). As a reference, the average incidence rate at state level was equal to 12 cases per 100,000 people per day from March 1st to October 31st 2020. Extrapolating to the entire US population, a three-day period with one additional °C of temperature leads to 868 fewer cases one week later. A three-day period with one standard deviation of temperature higher than the monthly mean in each state causes approximately 2,456 fewer cases at national level. A week with temperature consistently one standard deviation above the monthly average causes 17,193 fewer cases at national level seven days later. By applying a mortality rate equal to 1.7% of all positive cases, a three-day period with temperature one standard deviation above the state mean.

---

1 The average standard deviation of mean daily temperature over three days across all states and months was equal to 2.83 degree Celsius from March 1st to November 10th.
monthly average in all states leads to 42 fewer deaths at national level. Using a Value of Statistical Life equal to $10 million, the marginal benefit of one day with temperature one standard deviation larger than the monthly average in the whole country is equal to $420 million. A week with temperature constantly above the state average for that month leads to 290 fewer deaths and to a reduced economic cost equal to $2.9 billion.

Precipitation and cloud coverage do not significantly affect the incidence rate in our preferred specification. Relative humidity positively and significantly affects the incidence rate, probably because high humidity favors indoor activities. Average incidence in neighboring states has a significant and positive effect on incidence in the state, suggesting spatial diffusion of the virus across state borders. However, the effect of incidence in neighboring states is small. One hundred additional cases per 100,000 people on average in neighboring states lead to approximately one additional case per 100,000 people one week later.

Using county data, we find similar results (column (1) in Table 2 in Extended Data). One additional °C of temperature reduces the incidence rate by 0.315 cases per 100,000 people. Rainfall, cloud coverage and, in this case, also relative humidity do not have a statistically significant effect on the incidence rate. The effect of cases in neighboring counties is positive and approximately six times larger than what we find when using state data. Averages for neighboring counties include the increase in cases for all counties that share a border with the initial county, including counties in other states. There are clearly more movements and contagions within smaller regional areas than between states.
Our results are robust to a number of alternative specifications. We prefer omitting lagged incidence among the explanatory variables because the lagged dependent variable includes possible effects of temperature and other control variables. Including lagged incidence reduces the magnitude of the temperature coefficient but not significantly (columns 2 in Table 1 and Table 2 in ED). We repeat the analysis using state-by-week fixed effects to control for possible infra-monthly events that may explain the incidence rate, such as public health regulations or other trends within each month. Results with both state and county data confirm that warmer days reduce the incidence rate (column 6 in Table 1 and Table 2 in ED). However, as we greatly limit the variation of temperature, the coefficients are smaller. The temperature coefficient remains significant at the 95% level using state data, but is significant only at the 90% level using county data. Finally, two placebo tests (see Methods) suggests that the relationship between temperature and the incidence rate that we find is not random (columns 7 and 8 in Table 1 and Table 2 in ED).

Our study is the first comprehensive analysis of the effect of temperature on the incidence rate of COVID19 that uses quasi-random variation in temperature as a source of identification. Most of the studies that use uncontrolled experiments use cross-sectional variation, do not control for time trends by region, use temperature only as a control variable in analysis of stay-at-home orders, or differ in many important ways from our analysis. Our study is the only one that exclusively focuses on the US and that uses all 48 contiguous US states plus the District of Columbia, and virtually all counties and county equivalents in these states (3,100 out of 3,109). Other studies pool data from different regions of the world, limit the analysis to about one fifth of US counties, merge some US counties and some Chinese cities, limit their analysis to
Bangladesh, and Wuhan (China). Overall, this literature finds that warming generally leads to fewer cases or rate of diffusion. Results cannot be easily compared due to differences in methods and in dependent variables, but the effect of temperature is usually not large.

**Heterogeneity over time and space**

The temperature coefficient in our main specification measures the average effect of temperature over time and states (counties). We test if the effect of temperature changes over time by interacting the temperature coefficient with month dummy variables. We find a more complex relationship between temperature and incidence (Figure 1 and coefficients in columns (3) of Table 1 and Table 2 in ED). With state data, the effect of temperature is not significantly different from zero until the beginning of autumn. In September, the effect of temperature becomes negative and in October is much larger when we assume a uniform effect over time. With county data, we find a similar but more pronounced pattern. In July, the effect of temperature is significantly greater than zero, implying that higher temperature relative to the state mean in July, leads to a higher incidence rate. By adding the lagged incidence rate, we find similar results but the temperature effects are smaller (column (4) of Table 1 and Table 2 in ED).

By interacting the temperature coefficient with state dummy variables, we find large spatial heterogeneity in how temperature affects the incidence rate of COVID19 (Figure 2). In some states, higher temperatures significantly increase the number of cases. For another group of states the effect of temperature is not significantly different from zero. However, for the majority of states, warmer temperatures significantly reduce the number of cases.
Notes: 95% confidence interval, standard errors clustered at state level for both state and county data. Red lines plot the marginal effect estimated with our base model. The black lines indicate the marginal effect estimates by interacting temperature with months. Linear interpolation between the point estimates.

**Figure 1.** The impact of 1 °C warmer day on COVID19 incidence rate.

Notes: In lighter shades states for which the temperature effect is not significantly different from zero at the 95% level. The upper panel uses state data; the bottom panel uses county data.

**Figure 2.** The marginal effect of temperature on the incidence rate, by state.
Non-linear effect of temperature

We suspect that average temperature in each of these states plays a role in explaining the differences across space of temperature marginal effects. Explaining these differences is beyond the scope of our analysis because our panel data method cannot identify the effect of temperature levels. The effect of temperature levels can be studied using a cross-sectional analysis that controls many potential confounders. Here, we show a simple correlation analysis that reveals that average temperature and the marginal effect of temperature are positively correlated, especially when we use county data. This result may be explained by a non-linear effect of temperature on the incidence rate. When temperature is relatively low, a day warmer than average encourages outdoor activities. When temperature is relatively high, a day warmer than average encourages indoor activities.

We find additional evidence of non-linear temperature effects when we estimate the size and sign of the deviation from average monthly temperature by state (Figure 4). Largest deviations from the average have the largest impact on the incidence rate. Days that are more than 6 °C colder than average increase the incidence rate by approximately 4 units per 100,000 people. The effect of colder days quickly diminishes but it remains positive and significant. Similarly, warmer days that are close to the average have a small but significant negative effect on the incidence rate and days that are more than 6 °C warmer than average reduce the incidence rate by approximately 3 cases per 100,000 people. We find similar results using county data (Figure 1 in ED).
Figure 3. Correlation between the marginal effect of temperature on the incidence rate and average temperature, by state.

Notes: The Figure displays coefficients of dummy variables that are equal to 1 if the deviation of lagged average temperature with respect to the monthly average in each state is equal to the corresponding temperature interval measured on the horizontal axis. 95% confidence interval using standard errors clustered at state level.

Figure 4. The effect of deviations from average temperature using state data.
The effect of weather on time spent home

We advance the hypothesis that the effect of temperature on the incidence rate is explained by the time spent indoors versus the time spent outdoor. However, our main specification of the model is not able to separate the environmental (direct) and behavioral (indirect) effects of weather on the virus. To shed light on the relative importance of direct and indirect effects, we use data on time spent home and on other indicators of mobility from Safegraph. This data relies on anonymous GPS coordinates from mobile phones collected from millions of users with very high frequency. From repeated observations, it is possible to infer whether mobile phone users are at home, for how long and how far users travel away from their residence.

We estimate our base model by introducing several of these behavioral variables, jointly and independently, with lags of different lengths, and we find that they have ambiguous effects on the incidence rate. In some cases, the coefficients have signs opposite to what one would expect. For example, time spent at home positively affects the incidence rate. We believe that behavioral variables and variables that control for social-distancing are very likely endogenous and should be omitted from the model. The positive sign of the time spent at home variable is an indication of possible reverse causality.

We then test the hypothesis that weather explains stay-at-home behavior by regressing the average time spent at home in each state (or county) on weather variables and on the local and neighboring incidence rate, with weekday and state-by-month fixed effects. We find that all the four weather variables are highly significant (Table 3 in ED). Days warmer than average push people to spend more time outside their homes. Days with rainfall and cloud coverage higher than average push
people to spend more time at home. Days with higher humidity also lead to more time at home, probably because humidity is considered a nuisance. On average, people spend the least time at home during the three days that precede Saturdays (Wednesdays, Thursdays and Fridays), and the most during the three days that precede Tuesdays (Saturdays, Sundays and Mondays). This simple model explains 80% of the temporal variation of stay-at-home behavior within states and 44% of the temporal variation within counties. These results suggest that weather has a direct effect on the amount of time people spend at home vs outside their home and corroborates our hypothesis that weather has strong indirect effects on the incidence rate of COVID-19.

**Choice of lags, moving average and variable selection**

To validate the choice of lags for response and covariates, we repeat the analysis using lags from 5 to 9 days. We use the Akaike Information Criterion (AIC) to select the lag that leads to the most accurate model. Small AIC values correspond to more accurate models.

We begin with the lag selection for covariates. The AIC values for different lags along with the estimated temperature coefficient in the regression model corresponding to these lags are shown in Panel (a) of Figure 5. Seven-day lag results in the lowest AIC values for the regression model without an autoregressive term, which supports the choice in our preferred specification. We also allow lags for both the response and covariates to vary from 5 to 9 and we report the AIC values for each combination in Panel (b) of Figure 5. Using a seven-day lag for the response variable and a nine-day lag for the covariates leads to the lowest AIC value. The temperature coefficient in this case is equal to -0.213, slightly larger than when we use the same lag for both the response and the
covariates. An alternative approach to select the lag of the response variable is to run the regression model by constraining the lag of covariates to be 7, selected from the model without autoregressive term. We find that the optimal lag for the response variable using this approach is also 7 (Panel (c) of Figure 5). The temperature coefficients are always negative regardless of the lag chosen for the response and covariates. This shows the robustness of the results to the choice of lags.

To study the importance of the covariates in explaining the variation of the incidence rate, we perform variable selection using the regularization approach. Specifically, we apply the Elastic Net, which can deal with the multi-collinearity issue, to our regression models with lags 7. The estimated coefficients are reported in Table 4 in ED. The Elastic Net effectively shrinks the coefficients of the total cloud coverage and total precipitation to zero. Temperature and humidity are retained. Temperature is the most important weather covariate. We further use the Elastic Net method with the autoregressive term and for different lag combinations. The results are reported in ED and show that regardless of the choice of lags, the effect of the total precipitation and total cloud coverage shall be ignored, while the effect of temperature is always present.
Notes: Panel (a) uses different lags for all covariates but no lagged response variable. Panel (b) uses also the lagged dependent variable. The first number in parenthesis is the lag of the response variable, the second number is the lag of covariates. Panel (b) studies the optimal lag for the response variable when the lag of the covariates is 7. The figures display the temperature coefficient.

**Figure 5.** AIC criterion and temperature coefficients for different lags for response and covariates.
Conclusions

We provide the first quasi-experimental evidence of the effect of temperature on the incidence rate of COVID19 using state and county data in the US. We find that higher than average temperature leads to a lower incidence rate.

The effect of temperature on the incidence rate is relatively small, confirming the intuition of public health experts that environmental conditions alone have a marginal effect on the circulation of the novel coronavirus. However, we may underestimate the effect of different temperature across seasons and regions because we constrain our analysis to marginal random variations of temperature from state (or county) by month means to control for potential confounding effects. These random variations of temperature explain marginal changes in behavior after accounting for other seasonal changes and other regional characteristics and norms, including seasonal and regional climates. We capture the effect of random temperature changes on the probability that the play date, the birthday party, the restaurant dinner, the family reunion, will be indoors instead of being outdoors. We may thus provide only a lower bound to the overall effect of temperature on the incidence rate of COVID19.
Methods

Data
We compile a panel of COVID19 cases per 100,000 people (incidence rate), daily mean temperature and other control variables for the 48 contiguous US states and the District of Columbia from March 1st to October 31st, 2020. We use daily cases of COVID19 from the Atlantic COVID Tracking Project for states and The New York Times for counties, hourly weather from ERA5 reanalysis was aggregated to daily weather, and mobile phone data from SafeGraph. Population data is a 2019 estimate from the United States Census Bureau. COVID19 cases and mobile phone data is matched to counties and states. Dates with irregularities in cases reported, such as a negative increase in cases, were dropped from the analysis. Mobile phone data is provided on a census block group level and aggregated to the county level, and then aggregated once again to the state level. Weather data is matched to counties and states by taking the area-weighted average of ERA5 grid cell values. The average incidence rate in neighboring states and counties is calculated by averaging the incidence rate in all the neighboring states and counties.

Model specification
In our preferred specification we estimate the effect of temperature on the incidence rate of COVID19 using a panel model with fixed effects. For any variable, $x_{s,t}$ denotes the value taken by the variable in day $t$ in state $s$. $\bar{x}_{i,s}$ denotes the three-day moving average filter applied to variable $x_{s,t}$:

$$\bar{x}_{s,t} = \frac{1}{3} \sum_{r=1}^{3} x_{s,t-r}.$$ 

We estimate the following model for:
\[ \bar{y}_{s,t} = \alpha + \beta \bar{T}_{s,t-7} + \gamma \bar{X}_{s,t-7} + \gamma_w + \lambda_{s,m} + \epsilon_{s,t} \]

Where \( \bar{y}_{s,t} \) is the three-day moving average of positive tests per 100,000 people (the incidence rate), \( \bar{T}_{s,t-7} \) is average daily temperature with a three-day moving average filter and a 7-day lag, \( y \) is a \((1 \times J)\) vector of parameters and \( X_{s,t} \) is a \((J \times 1)\) vector of time-variant control variables to which we apply the three-day moving average filter, with a 7-day lag. In our preferred specification, the control variables include rainfall (cm/day), relative humidity (%), cloud coverage (%) and the incidence rate in neighboring states or counties. \( \gamma_w \) is a fixed effect that controls for the day of the week in which the incidence rate is measured. \( \lambda_{s,m} \) is a fixed effect that varies by state and month. \( \epsilon_{s,t} \) is the error component. The model identifies the causal effect of temperature on the incidence rate if it is correctly specified and if the error term is exogenous. The state-by-month fixed effects capture unobserved heterogeneity at state level and by month. This allows to control for all local effects that do not change over the time of our panel – e.g. geography, climate, demographics, governance, religion, and ethnicity – and for all things that change over time within each state – e.g. policy responses to COVID19, seasonal changes in behavior, and seasonal changes in climate. This restrictive conditions almost completely remove any trend in the dependent variable. The weekday fixed effects control for the observed weekly pattern in testing and in the release of test results. Our method is equivalent to subtracting state, month and weekday averages and then estimating the model using OLS. In robustness tests we use a state-by-week fixed effect. We cluster standard errors at state level to account for possible correlation within states.

Similarly, for counties we estimate:

\[ \bar{y}_{c,t} = \alpha + \beta \bar{T}_{c,t-7} + \gamma \bar{X}_{c,t-7} + \gamma_w + \lambda_{s,m} + \eta_c + \epsilon_{c,t}. \]
Where $c$ denotes a county and $\eta_c$ is a county fixed effect that removes all the unobserved, time invariant, heterogeneity at county level. We use state-by-month fixed effects to remove broad state-wide heterogeneity and time trends at state level.

Models that include the lagged dependent variable are specified as follows:

$$\bar{y}_{s,t} = \alpha + \beta T_{s,t-7} + \gamma X_{s,t-7} + \gamma_w + \lambda_{s,m} + \delta \bar{y}_{s,t-7} + \epsilon_{s,t}.$$ 

To estimate the effect of temperature separately by month, we interact the temperature coefficient with month fixed effects:

$$\bar{y}_{s,t} = \alpha + \sum_{m=3}^{10} \beta_m T_{s,t-7} + \gamma X_{s,t-7} + \gamma_w + \lambda_{s,m} + \epsilon_{s,t}.$$ 

Similarly, to estimate the effect of temperature by state, we interact temperature with a state fixed effect:

$$\bar{y}_{s,t} = \alpha + \sum_{s=1}^{49} \beta_s T_{s,t-7} + \gamma X_{s,t-7} + \gamma_w + \lambda_{s,m} + \epsilon_{s,t}.$$ 

We proceed analogously when using county data.

To estimate the effect of deviations from average temperature by size of the deviation, we interact deviations of temperature from the mean by state, month and day of the week with dummy variables for different sizes of the deviation below and above the mean.

**Placebo tests**

To rule out the possibility that the relationship between temperature and the incidence rate is purely random, we run two placebo tests. In the first test, we randomly shuffle the dependent variable (incidence rate) by state (or county), so that daily observations of COVID19 are randomly assigned
to daily observations of all the independent variables. In the second test, we randomly shuffle temperature by state (or county), so that we randomly assign temperature to each day count of COVID19 cases and to the other covariates. We expect that coefficients of the independent variables are not statistically significant in the first placebo test and we expect that the second test reveals that the temperature coefficient is not statistically significant.

**Lag and variable selection**

To determine which lag day $k$ is most appropriate, we utilize the AIC as the model selection criterion. AIC is built upon information theory, which is a trade-off between the goodness-of-fit and model simplicity. In other words, a model gets rewarded if it fits the data well, whereas it gets penalized for being complex. The model with lower AIC scores would be preferred. In order to calculate the AIC score, we define the AIC for our regression model as follows:

$$AIC = -\frac{2}{N \times LL} + 2 \times \frac{K}{N'},$$

where $N$ is the number of data points, $LL$ is the log-likelihood of the training data and $K$ is the number of parameters used in the model. The pairs of lag values that result in lower AIC scores are used in our study.

For variable selection and removing the effect unimportant covariates, we use Elastic Net, which is comprised of an ordinary least square loss function and a penalty term defined by a linear combination of $L_1$ and $L_2$ norms of the coefficients. The $L_1$ norm can help remove unimportant variables and the $L_2$ norm deals with the multicollinearity issue that might produce unreliable results due to high standard errors of the estimated coefficients. Specifically, the objective function for the regression problem using Elastic Net can be mathematically formalized as such:
\[
\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|eta\|_1 + \lambda_2 \|eta\|_2^2,
\]

where \(\lambda_1\) and \(\lambda_2\) are the tuning parameters. In order to choose the right tuning parameters, we perform five-fold cross validation.
Data availability

All data generated or analyzed during this study is available from the corresponding author on reasonable request. Publication in a data repository will be considered in case the paper is accepted for publication. Some restrictions apply to SafeGraph data on mobile phone use, which are not publicly available and were used under license for the current study. Data are however easily available at no cost from SafeGraph after submitting a request. If the paper is accepted for publication, we will check with SafeGraph about the possibility of republishing their data.

Code availability

All the code used to prepare and analyze the data during this study is available from the corresponding author. If the paper is accepted for publication, distribution of the code in a data repository will be considered.
References

1. Morris, D. H. et al. The effect of temperature and humidity on the stability of SARS-CoV-2 and other enveloped viruses. bioRxiv (2020).
2. Altizer, S. et al. Seasonality and the dynamics of infectious diseases. Ecology Letters 9, 467-484, doi:https://doi.org/10.1111/j.1461-0248.2005.00879.x (2006).
3. The Atlantic. The Covid Tracking Project, <https://covidtracking.com/data/download> (2020).
4. The New York Times. Coronavirus in the U.S.: Latest Map and Case Count, <https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html> (2020).
5. Hersbach, H. et al. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999-2049 (2020).
6. SafeGraph. Social Distancing Metrics, <https://docs.safegraph.com/docs/social-distancing-metrics> (2020).
7. World Health Organization. Coronavirus disease (COVID-19), <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19> (2020).
8. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases 20, 533-534 (2020).
9. Wu, Y. et al. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Science of the Total Environment, 139051 (2020).
10. Wang, J., Tang, K., Feng, K. & Lv, W. High temperature and high humidity reduce the transmission of COVID-19. Available at SSRN 3551767 (2020).
11. Ma, Y. et al. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of The Total Environment 724, 138226, doi:https://doi.org/10.1016/j.scitotenv.2020.138226 (2020).
12. Haque, S. E. & Rahman, M. Association between temperature, humidity, and COVID-19 outbreaks in Bangladesh. Environmental Science & Policy 114, 253-255, doi:https://doi.org/10.1016/j.envsci.2020.08.012 (2020).
13. Li, A. Y. et al. Multivariate analysis of black race and environmental temperature on COVID-19 in the US. The American journal of the medical sciences 360, 348-356 (2020).
14. Palialol, B., Pereda, P. & Azzoni, C. Does weather influence COVID-19 transmission? Regional Science Policy & Practice 12, 981-1004 (2020).
15. Allcott, H. et al. What Explains Temporal and Geographic Variation in the Early US Coronavirus Pandemic? NBER Working Paper (2020).
16. Araujo, M. B. & Naimi, B. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. medRxiv (2020).
17. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67, 301-320 (2005).
18. United States Census Bureau. Total Population. (2019).
Acknowledgements

We are grateful to the following organizations for free access to their data: NOAA/OAR/ESRL PSL, Boulder, Colorado, USA for the NCEP North American Reanalysis; SafeGraph, for mobile phone data; The Atlantic COVID Tracking Project and The New York Times for COVID19 data. Emanuele Massetti and Eric Dobbie acknowledge financial support from DILAC, Ivan Allen College of Liberal Arts, Georgia Tech.

Author Contributions

E.M. and K.P. conceptualized the analysis. E.D. collected and processed data. E.M., E.D. and W.Y. performed the statistical analysis. The authors contributed equally to the analysis of the results. E.M. wrote the first draft of the paper. All authors have reviewed and edited the paper.

Competing Interests

The authors declare no competing interests.

Materials & Correspondence

Emanuele Massetti, School of Public Policy, Georgia Institute of Technology, Atlanta, GA, United States, emanuele.massetti@gatech.edu

Additional Information

Supplementary Information is available for this paper.
## Extended Data

|                  | (1)         | (2)         | (3)         | (4)         | (5)         | (6)         | (7)         | (8)         |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Temperature (C)  | -0.263***   | -0.195***   | -0.081**    | 0.074       | -0.005      |             |             |             |
|                  | (0.061)     | (0.032)     | (0.026)     | (0.044)     | (0.008)     |             |             |             |
| Temperature by Month (C) |            |             |             |             |             |             |             |             |
| March            | 0.016       | -0.026      |             |             |             |             |             |             |
|                  | (0.046)     | (0.027)     |             |             |             |             |             |             |
| April            | 0.015       | -0.005      |             |             |             |             |             |             |
|                  | (0.042)     | (0.043)     |             |             |             |             |             |             |
| May              | -0.046      | -0.067      |             |             |             |             |             |             |
|                  | (0.050)     | (0.048)     |             |             |             |             |             |             |
| June             | 0.001       | 0.034       |             |             |             |             |             |             |
|                  | (0.050)     | (0.035)     |             |             |             |             |             |             |
| July             | 0.233       | 0.042       |             |             |             |             |             |             |
|                  | (0.116)     | (0.088)     |             |             |             |             |             |             |
| August           | -0.019      | -0.075      |             |             |             |             |             |             |
|                  | (0.132)     | (0.107)     |             |             |             |             |             |             |
| September        | -0.243**    | -0.223***   |             |             |             |             |             |             |
|                  | (0.077)     | (0.054)     |             |             |             |             |             |             |
| October          | -1.106**    | -0.705***   |             |             |             |             |             |             |
|                  | (0.221)     | (0.125)     |             |             |             |             |             |             |
| Rainfall (cm/day)| -0.503*     | -0.316      | -0.216      | -0.160      | -0.274      | -0.309      | 0.159       | -0.342      |
|                  | (0.248)     | (0.187)     | (0.192)     | (0.172)     | (0.212)     | (0.184)     | (0.603)     | (0.235)     |
| Relative Humidity (%) | 0.037*    | 0.033**    | 0.045**    | 0.039***    | 0.022      | 0.007      | 0.037      | 0.041*      |
|                  | (0.015)     | (0.011)     | (0.013)     | (0.011)     | (0.013)     | (0.011)     | (0.033)     | (0.017)     |
| Cloud Coverage (%) | -0.039    | -0.121      | -0.930      | -0.673      | -0.359      | 0.968*      | -1.170      | -0.254      |
|                  | (0.511)     | (0.387)     | (0.587)     | (0.429)     | (0.498)     | (0.439)     | (1.470)     | (0.537)     |
| Incidence Rate in Neighbors | 0.007*** | 0.002***    | 0.007***    | 0.002***    | 0.007***    | 0.004***    | -0.001      | 0.007***    |
|                  | (0.001)     | (0.001)     | (0.001)     | (0.001)     | (0.001)     | (0.001)     | (0.000)     | (0.001)     |
| Lagged Incidence  | 0.637***    | 0.600***    |             |             |             |             |             |             |
|                  | (0.055)     | (0.045)     |             |             |             |             |             |             |
| N                | 11,445      | 11,336      | 11,445      | 11,336      | 11,445      | 11,445      | 9,121       | 11,409      |
| State by Month FE | Y           | Y           | Y           | Y           | Y           | N           | Y           | Y           |
| State by Week FE | N           | N           | N           | N           | N           | Y           | N           | N           |
| Weekday FE       | Y           | Y           | Y           | Y           | Y           | Y           | Y           | Y           |
| Placebo Test     | --          | --          | --          | --          | --          | --          | Y           | Y           |

Notes: Significance levels: * p<0.05; ** p<0.01; *** p<0.001. Standard errors clustered at state level. (1) is our preferred specification. (2) adds lagged incidence rate. (3) and (4) interact temperature with month fixed effects, without and with the lagged incidence rate. (5) reports coefficients other than temperature when the temperature variable is interacted with state fixed effects. (6) uses state by week fixed effects. (7) and (8) are two placebo tests: (7) randomly shuffles the dependent variable by state while (8) shuffles temperature by state.

Table 1. Summary of results using state data.
|                           | (1)          | (2)          | (3)          | (4)          | (5)          | (6)          | (7)          | (8)          |
|---------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Temperature (C)           | -0.315***    | -0.274***    | -0.150*      | 0.017        | -0.004       |
|                           | (0.073)      | (0.065)      | (0.061)      | (0.016)      | (0.003)      |
| Temp. by Month (C)        |              |              |              |              |              |
| March                     | -0.264***    | -0.122***    |              |              |              |
|                           | (0.066)      | (0.099)      |              |              |              |
| April                     | -0.075       | -0.056       |              |              |              |
|                           | (0.056)      | (0.053)      |              |              |              |
| May                       | -0.111       | -0.101       |              |              |              |
|                           | (0.066)      | (0.057)      |              |              |              |
| June                      | -0.029       | -0.017       |              |              |              |
|                           | (0.056)      | (0.048)      |              |              |              |
| July                      | 0.501*       | 0.378*       |              |              |              |
|                           | (0.197)      | (0.167)      |              |              |              |
| August                    | 0.087        | 0.047        |              |              |              |
|                           | (0.086)      | (0.079)      |              |              |              |
| September                 | -0.295**     | -0.283***    |              |              |              |
|                           | (0.099)      | (0.090)      |              |              |              |
| October                   | -1.091***    | -0.894***    |              |              |              |
|                           | (0.173)      | (0.157)      |              |              |              |
| Rainfall (cm/day)         | -0.087       | -0.025       | 0.131        | 0.151        | 0.244        | 0.121        | 0.092        | 0.019        |
|                           | (0.270)      | (0.265)      | (0.235)      | (0.234)      | (0.294)      | (0.262)      | (0.120)      | (0.273)      |
| Relative Humidity (%)     | 0.010        | 0.015        | 0.025        | 0.027        | -0.006       | -0.008       | -0.002       | 0.016        |
|                           | (0.016)      | (0.016)      | (0.014)      | (0.015)      | (0.014)      | (0.015)      | (0.006)      | (0.016)      |
| Cloud Coverage (%)        | -0.478       | -0.600       | -1.256*      | -1.246*      | -1.028       | 0.870        | -0.159       | -0.366       |
|                           | (0.658)      | (0.679)      | (0.590)      | (0.592)      | (0.623)      | (0.500)      | (0.300)      | (0.712)      |
| Inc. Rate in Neighbors    | 0.048**      | 0.031**      | 0.045**      | 0.029**      | 0.048***     | 0.036*       | 0.001        | 0.049***     |
|                           | (0.014)      | (0.010)      | (0.014)      | (0.010)      | (0.018)      | (0.014)      | (0.001)      | (0.014)      |
| Lagged Incidence          |              |              |              | 0.236***     | 0.233***     |
|                           |              |              |              | (0.059)      | (0.058)      |
| N                         | 639,680      | 623,069      | 639,680      | 623,069      | 639,680      | 639,680      | 474,024      | 637,652      |
| County FE                 | Y            | Y            | Y            | Y            | Y            | Y            | Y            | Y            |
| State by Month FE         | Y            | Y            | Y            | Y            | Y            | N            | Y            | Y            |
| State by Week FE          | N            | N            | N            | N            | N            | Y            | N            | N            |
| Weekday FE                | Y            | Y            | Y            | Y            | Y            | Y            | Y            | Y            |
| County by Month FE        | --           | --           | --           | --           | --           | --           | --           | --           |
| Placebo Test              | --           | --           | --           | --           | --           | --           | Y            | Y            |

Notes: Significance levels: * p<0.05; ** p<0.01; *** p<0.001. Standard errors clustered at state level. (1) is our preferred specification. (2) adds lagged incidence rate. (3) and (4) interact temperature with month fixed effects, without and with the lagged incidence rate. (5) reports coefficients other than temperature when the temperature variable is interacted with state fixed effects. (6) uses count by week fixed effects. (7) and (8) are two placebo tests: (7) randomly shuffles the dependent variable by county while (8) shuffles temperature by county.

Table 2. Summary of results using county data.
| Variable                          | State   | County  |
|----------------------------------|---------|---------|
| Temperature (°C)                 | -1.747*** | -1.508*** |
|                                  | (0.080)  | (0.057)  |
| Rainfall (cm/day)                | 3.093*** | 2.752*** |
|                                  | (0.727)  | (0.419)  |
| Relative Humidity (%)            | -0.278*** | -0.168*** |
|                                  | (0.037)  | (0.022)  |
| Cloud Coverage (%)               | 11.683*** | 6.157*** |
|                                  | (1.849)  | (1.265)  |
| Incidence Rate in Neighbors      | 0       | 0.004   |
|                                  | (0.002)  | (0.009)  |
| Incidence in State/County        | 0.113   | 0.050*** |
|                                  | (0.074)  | (0.007)  |
| Weekday                          |         |         |
| Monday                           | 13.899*** | 16.291*** |
|                                  | (1.293)  | (1.146)  |
| Tuesday                          | 18.966*** | 25.888*** |
|                                  | (1.215)  | (1.318)  |
| Wednesday                        | 15.583*** | 23.125*** |
|                                  | (0.726)  | (0.962)  |
| Thursday                         | 18.966*** | 25.888*** |
|                                  | (0.236)  | (0.218)  |
| Friday                           | (omitted)|         |
| Saturday                         | -4.090*** | -7.394*** |
|                                  | (0.277)  | (0.262)  |
| Sunday                           | -0.788   | -5.511*** |
|                                  | (0.794)  | (0.553)  |
| State by Month FE                | Y       | Y       |
| N                                | 8,183    | 507,311  |
| Adj R² (within variation)        | 0.798    | 0.439    |

Notes: The dependent variable is average time spent home, measured in minutes. All variables smoothed using a three-day moving average filter. Weekday fixed effects control for the effect of the three preceding days. Significance levels: * p<0.05; ** p<0.01; *** p<0.001. Standard errors are clustered at state level.

Table 3. The effect of weather on time spent home.
|                      | (1)   | (2)   | (3)   | (4)   | (5)   |
|----------------------|-------|-------|-------|-------|-------|
| Temperature (C)      | -0.263| -0.195| -0.190| -0.163| -0.128|
| Rainfall (cm/day)    | -0.502| -0.315| 0.024 | 0.000 | 0.022 |
| Relative Humidity (%)| 0.038 | 0.033 | 0.000 | 0.000 | 0.000 |
| Cloud Coverage (%)   | -0.071| -0.13  | 0.000 | 0.000 | 0.000 |
| Incidence Rate in Neighbors | 0.008 | 0.002 | 0.007 | 0.001 | 0.002 |
| Lagged Incidence     |       | 0.635 | 0.630 | 0.597 |
| Elastic Net          | N     | N     | Y     | Y     | Y     |
|                      |       |       |       |       | 11,435|
|                      |       |       |       |       | 11,326|
|                      |       |       |       |       | 11,435|
|                      |       |       |       |       | 11,326|
|                      |       |       |       |       | 11,326|

Notes: All models use state data. Columns (1) and (2) estimated using the fixed effect model as in Table 1. Columns (3) to (5) use Elastic Net for variable selection. Column (3): without lagged dependent variable. Columns (4) and (5) with lagged dependent variable, with lag 7 and 9 respectively. We do not report standard errors because they are not directly comparable across models.

**Table 4. Using Elastic Net for variables selection.**

![Figure 1: The effect of deviations from average temperature using county data.](image-url)
Detailed review of the literature

Wang et al. find that high temperature and high humidity reduce the transmission of COVID-19 using data from China and the US, but they use cross-section evidence without controlling for time trends. Allcott et al. study the effect of stay at home order using regression discontinuity analysis and control for seasonal temperature, but they do not find a significant effect and the estimation of temperature effects is not at the core of their analysis. Wu et. al use cross-section methods on data from 166 regions in the world, including the US and find that one additional °C is associated with a 3% reduction in daily new cases and a 1% increase in relative humidity is associated with a 0.85% reduction in daily cases. Palialol, Pereda and Azzoni use data from 416 regions of the world from February to April to study the effect of temperature on the total number of cases in a region, including fixed effects. They find one additional °C over a fifteen-day period reduces the number of cases by 9%. An important difference with our study is that our dependent variable is the incidence rate while they use the total number of cases. We prefer to work with the incidence rate because differences in population may distort the results. Another important difference is that we use region-by-month fixed effects to control for time trends that are specific to each region while they use region and month fixed effects separately, which constraints the time trend to be the same across all regions. This constraint on the time trend may be problematic when studying countries in both hemispheres. Araújo and Naimi use a model of ecological niches and machine learning methods to estimate how environmental factors affect the biological survival of the virus, but they do not control for socio-economic factors.