entropy

MDPI

Article

Understanding the Impact of Walkability, Population Density, and Population Size on COVID-19 Spread: A Pilot Study of the Early Contagion in the United States

Fernando T. Lima [1,2,*], Nathan C. Brown [3] and José P. Duarte [1]

[1] Stuckeman Center for Design Computing, The Pennsylvania State University, University Park, State College, PA 16802, USA; jxp400@psu.edu
[2] Faculty of Architecture and Urbanism, Universidade Federal de Juiz de Fora, Juiz de Fora, MG 36036-900, Brazil
[3] Department of Architectural Engineering, The Pennsylvania State University, University Park, State College, PA 16802, USA; ncb5048@psu.edu
* Correspondence: fernando.lima@arquitetura.ufjf.br

Abstract: The novel coronavirus disease 2019 (COVID-19) pandemic is an unprecedented global event that has been challenging governments, health systems, and communities worldwide. Available data from the first months indicated varying patterns of the spread of COVID-19 within American cities, when the spread was faster in high-density and walkable cities such as New York than in low-density and car-oriented cities such as Los Angeles. Subsequent containment efforts, underlying population characteristics, variants, and other factors likely affected the spread significantly. However, this work investigates the hypothesis that urban configuration and associated spatial use patterns directly impact how the disease spreads and infects a population. It follows work that has shown how the spatial configuration of urban spaces impacts the social behavior of people moving through those spaces. It addresses the first 60 days of contagion (before containment measures were widely adopted and had time to affect spread) in 93 urban counties in the United States, considering population size, population density, walkability, here evaluated through walkscore, an indicator that measures the density of amenities, and, therefore, opportunities for population mixing, and the number of confirmed cases and deaths. Our findings indicate correlations between walkability, population density, and COVID-19 spreading patterns but no clear correlation between population size and the number of cases or deaths per 100 k habitants. Although virus spread beyond these initial cases may provide additional data for analysis, this study is an initial step in understanding the relationship between COVID-19 and urban configuration.

Keywords: COVID-19; COVID-19 spread; walkability; population density; population size

Citation: Lima, F.T.; Brown, N.C.; Duarte, J.P. Understanding the Impact of Walkability, Population Density, and Population Size on COVID-19 Spread: A Pilot Study of the Early Contagion in the United States. *Entropy* **2021**, *23*, 1512. https://doi.org/10.3390/e23111512

Academic Editors: José A. Tenreiro Machado and Dimitri Volchenkov

Received: 19 September 2021
Accepted: 10 November 2021
Published: 14 November 2021

1. Introduction

The novel coronavirus pandemic has significantly changed the way people interact with each other and with urban space. As there are no fully immunized cities yet, there is scientific consensus on the importance of adopting social distancing strategies to control the spread of COVID-19, especially in areas of high transmission rates or low vaccination rates [1–4]. Over time, many significant, increasing, and competing issues have arisen from implementing aggressive social distancing measures versus preserving socioeconomic activity. In this regard, there is a knowledge gap regarding how urban environments' characteristics impact the spread of COVID-19 and infectious diseases in general. There is also a lack of, and an urgent demand for, data-driven approaches that can support decision-making related to these issues for different urban scenarios. The COVID-19 pandemic and all the complex data it generates point to the simple fact that contact leads to infection, and more face-to-face interactions are likely to increase transmission [1,3,4].

In this sense, cities are the stage where contact between people, and therefore infection, is more likely to occur. Although individuals and community behaviors such as how different groups choose to mask, stay home, get vaccinated and take other preventive measures also influence contagion spread, the available data indicates varying patterns of the spread of COVID-19 within American cities, especially in the first months when the contagion was faster in high-density and more walkable counties such as New York than in low-density and car-oriented counties such as Los Angeles (Figure 1). Thus, if an understanding is developed as to how urban features are correlated to different modes of social interaction and, consequently, to different patterns of COVID-19 spread, it will assist the identification of appropriate strategies to contain and mitigate the infection.

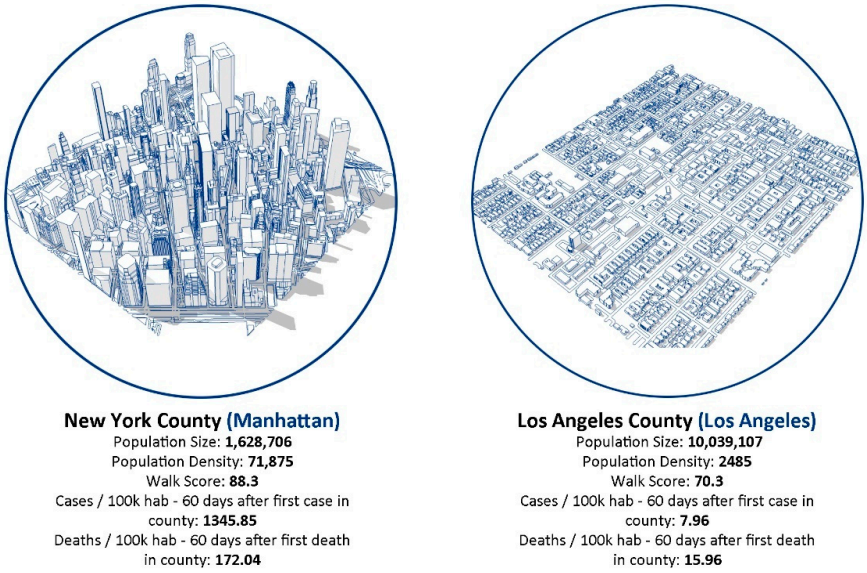

Figure 1. Basic urban features and COVID-19 dissemination data 60 days after the first case and the first death, according to USA Facts Database, in two counties addressed in this study: New York County and Los Angeles County.

From a city-as-a-network (or as a network of networks) perspective, approached by several researchers [5–9], several localities made decisions to interrupt the transmission of COVID-19 through total lockdowns, which is akin to dismantling the entire network [9]. It is thus imperative to understand how such networks function and the effect of their links, hubs, clusters, gates, and so on. To obtain this understanding, it is crucial to investigate the role of urban features in infection dynamics.

This paper is part of a larger study aimed at the study of the correlation between urban form and Covid 19 propagation patterns. The larger study includes features related to population, spatial configuration, use patterns, and climate on the one hand, and features related to disease and control features on the other hand. The paper presents an exploratory study based on a set of regression analyses. The primary goal is to address an initial set of variables in a proof-of-concept experiment that seeks to preliminarily verify our hypothesis that certain urban features and associated spatial use patterns correlate with the spread of COVID-19. If an understanding is developed in this regard, it can partially explain and predict the future spread while identifying appropriate strategies to contain and mitigate the infection. The paper is focused on variables that are regulated in urban design tasks and for which accurate data is readily available. Namely, this paper thus verifies possible correlations between the COVID-19 spread in the United States and the following urban features: (i) walkability; (ii) population density; (iii) population size, and; (iv) the number of days in stay-at-home order for each location. To identify the influence of walkability, population density, and population size on the dissemination of COVID-19 in the United States, we have considered the 93 urban American counties (with population sizes from

200 k to 10 M) for which walkability data (Walk Score) was available and a 60-day time-lapse from the first case confirmation and death dates in each one of them according to USAFacts Database. This paper is structured in the following sequence: (i) short review on the COVID-19 pandemic and urban features (walkability, population density, and population size); (ii) description of the performed regression analyses; (iii) presentation and discussion of the results and; (vi) final remarks about the study and identification of intended future developments.

2. Background: Urban Features and Infectious Diseases Spread

According to [10], urban areas are the ground zero of the COVID-19 pandemic, responsible for 90 percent of reported cases before and during April 2020. The United States, in turn, was the number one country in the world in terms of COVID-19 cases and deaths as of 15 March 2021. From 21 January 2020 to 28 April 2020, 1,006,417 confirmed cases and 57,433 deaths were reported in the U.S.

There is a knowledge gap regarding how the configuration of urban environments impacts the spread of infectious diseases. As mentioned above, this work is an initial step in a broader research agenda that seeks to characterize and codify the relationship between urban features, social interaction patterns, and COVID-19 transmission, particularly in the context of American cities. This paper presents early results that, combined with others to come, can offer directives for designing alternative containment strategies for different urban contexts, from compact and walkable neighborhoods to sparse and car-oriented districts, and from the scale of ZIP code areas to counties. We intend to provide guidelines for interventions in existing cities to make them more resilient to infectious diseases and the future design of resilient cities. While the datasets and corresponding knowledge are specific to COVID-19 in the United States, our established methodology could be extended to predict the spread of future epidemics in other urban areas.

2.1. Walkability

According to several authors [11–14], a walkable urban area or an urban area that follows walkability's principles considers pedestrians the highest priority, seeking greater urban life and promoting more socioeconomic interactions. According to this idea, walkability can be defined as a particular urban area's capability to connect housing and amenities from several categories (e.g., retail, food, education, entertainment, and recreation) through distances that can be traveled within walking distance. This means more people walking, cycling, staying in public spaces, interacting, and exchanging information, as well as social and cultural opportunities. Thus, higher walkabilities increase the likelihood of people meeting and interacting due to the higher density of amenities. In this sense, we hypothesize that walkability acts as a proxy for several social interaction-related features and that places with greater walkability promote higher social interaction levels and, therefore, higher contagion rates of certain contagious diseases, such as COVID-19. In other words, we advocate that when walkability (as understood as a metric for the density of services) and population density increase, the likelihood of people meeting in places such as transport stations, public facilities, common entries, and elevators also increases. This effect might have been more apparent before other significant factors came into play. For example, masks were not recommended in the U.S. until 3 April 2020, and vaccines were not widely available until the following year. Accordingly, the work of [15] shows that, in Italy, the highest spread rates occurred in areas with commercial hubs, close to the highest populated cities, and the most industrial area. Their results indicate how human mobility can affect the epidemic, identifying particular situations in which the health authorities can promptly intervene to control the spread of the disease. Urban features in turn affect human mobility, and their influence is worth studying as well.

Several studies addressed ways of measuring the walkability of a particular location. For instance, the works of [16–19] consider the structure of street networks and their number of intersections, among others. However, in this research, we adopted the walk

score index [20,21] for the following reasons: it is one of the most accessible walkability metrics (there are a lot of data available regarding the walk score of streets, neighborhoods, and cities in the U.S.); it was considered a reliable and valid measure of estimating walkable access to amenities; and walk score may be a convenient and inexpensive option for researchers interested in exploring the relationship between access to walkable amenities and health behaviors [22].

Walk score is an algorithmically obtained index for measuring an urban area's walkability by assigning a score to a location based on its distance to various nearby services. The amenities considered by walk score can be divided into five categories: educational (e.g., schools), retail (e.g., grocery, drug, convenience, and bookstores), food (e.g., restaurants), recreational (e.g., parks and gyms), and entertainment (e.g., movie theaters). The algorithm calculates the distance to the closest of each of the five amenities categories. The results are normalized to a 0 to 100 scale, considering 0 as the lowest walkability (car dependent) and 100 as the highest (most walkable). For example, in relation to a particular locality, if one of the five amenities is within a 0.4 km (5 min walk) radius from the input location, then the maximum number of points, 100, is assigned to it. The number of points decreases as the distance increases to 1.6 km (30 min walk), and no points are awarded for locations amenities farther than 1.6 km. For instance, New York County and San Francisco County have high Walk Score indexes (88.3 and 87.4, respectively), while Chesapeake (Virginia) and Cumberland County (North Carolina) have extremely low walk score indexes (21 and 21.4, respectively).

2.2. Population Density

55% of the world's population currently lives in urban areas, and this proportion is expected to increase to 68% by 2050 [23]. With people living in denser conditions, more interactions between individuals and disease transmission tend to occur more easily. As population density is an important urban feature that increases contact and, consequently, infection between people, several authors have studied the effect of population density on epidemic outbreaks in different contexts [24–26]. Still, the idea of high density of both population and buildings in urban areas is defended by several authors [12,27–29]. In the United States, population density is very heterogeneously distributed. For instance, New York County, Kings County, and Bronx County (all in New York) shelter, respectively, 71,876, 37,233, and 34,058 people per square mile. Washoe County (Nevada), Webb County (Texas), and San Bernardino Country (California) shelter, respectively, 74, 82, and 108 people per square mile.

2.3. Population Size

In addition to density and walkability, various socioeconomic interactions play an essential role in the dynamics of urban areas. As the overall size of a city is a critical aspect in defining social and economic life, it is also a relevant data point. Schläpfer et al. [30] advocate that different socioeconomic quantities increase superlinearly with city size and that this logic applies to almost all urban aspects, including the creation of new inventions and the prevalence of certain contagious diseases, for instance. At the same time, [31] state that the COVID-19 attack rate increases with city size and, in the absence of adequate controls, larger cities (and counties, as we assume) are expected to have more extensive epidemics than smaller ones. In the context of the United States and following this idea, Los Angeles County, California (10,039,107 inhabitants), Cook County, Illinois (5,150,233 inhabitants), and Harris County, Texas (4,713,325 inhabitants) would have the highest COVID-19 prevalence. Considering various hypotheses regarding relationships between different urban features, the population size should be compared to other factors.

2.4. Related Work

The recent COVID-19 pandemic stimulated the emergence of studies on the impact of population, spatial, and climatic features on the propagation of COVID-19 [32–34].

However, these studies are partial since they focus on just one or a few urban aspects. In addition, they are mainly focused on Chinese cities. On the other hand, Carozzi [35] states that density has affected the outbreak's timing in American counties, with denser locations more likely to have a stronger outbreak. In turn, Oishi, Cha, and Schimmack [36] analyzed the role of walkability, wealth, and race in New York City, finding that walkability was negatively related to the number of COVID-19 cases and deaths. However, at the same time, the same authors identified that areas with a higher presence of certain ethnicities, median age, and occupants per room were more likely also to have higher COVID-19 cases and deaths. Dasgupta et al. [37] and Rocha et al. [38] address the role of socioeconomic vulnerability in the U.S. counties and Brazil, respectively. However, there is still a knowledge gap regarding how the characteristics of urban environments impact the spread of COVID-19 and infectious diseases in general. This work aims to contribute to bridging this gap by presenting an approach that seeks to find the relationship between urban features, social interaction patterns, and COVID-19 transmission, particularly in the context of American counties.

3. Method

3.1. Data

To verify whether there are correlations between certain urban features (walkability, population density, population size) and COVID-19 spreading patterns in urban areas, this work focus on county-level data, instead of city-level data, for two reasons: county-level data allow us to consider larger areas and more significant populations, but at a level of granularity that distinguishes between various townships (from big cities to surrounding small towns); and most of the available data on COVID-19 is organized at the county level. Thus, in addition to its practicality, we believe that addressing county-level data can provide more comprehensive information about the role of urban networks, enabling broader conclusions and increased freedom of analysis.

Instead of addressing a single and national timeline, considering the day of the first case in the United States as day one for all counties, we decided to study how the disease spread in diverse locations to identify how different urban features and associated urban patterns correlated. Our logic considers each county, regardless of their particularities, as a preliminary token to understand the whole country. To overcome potential bias in the timing of the disease's onset across locations, we addressed the time-adjusted number of known cases and deaths per 100 k inhabitants in the studied counties. To this end, we considered two time-lapses for each county: 60 days after the first case (when addressing cases per 100 k hab) and 60 days after the first death (when addressing deaths per 100 k hab). The goal was to observe the longest time span possible and, at the same time, focus on spread in initial stages (when we assume that containment measures had less time to exert influence), allowing us to identify the effects of urban features more clearly. We used the software Minitab to run the analysis and the Grasshopper plugin for Rhinoceros was used for plotting some of the visualizations. Preliminary model fitting studies, carried out to identify the most suitable time interval, indicated the first 60 days as the best choice. After the first 60 days, both for cases and deaths per 100 k habitants, our R-squared adj values started to decrease significantly, as depicted in Figure 2.

When considering how to assess spread, death tolls are a more accurate indicator of COVID-19 prevalence since data on COVID-19 cases might be reported with error due to variation in local testing strategy and capacity [35,39]. However, we decided to address and compare both known cases and death tolls, as different aspects of health systems and underlying populations can influence the latter [40]. The number of known cases and death tolls were obtained from [41]. These data were combined with walk score [21], the number of days under a state-issued stay-at-home order [42], the population size of the counties [43] (total number of counties' inhabitants), and the mean population density for each county [43] (total population/land area in sq miles). As Walk Score data were available on a city-level basis for 112 cities (from 200,000 to 10,000,000 inhabitants) and some cities

were in the same county, and some counties were in the same city, it was necessary to aggregate data from the previous 112 cities into a final sample of 93 counties. Our sampling (Figure 3) allowed us to approach a total population of 115,791,837 people (35.27% of the U.S. population), 645,764 COVID-19 known cases, and 52,946 deaths (considering time adjustments).

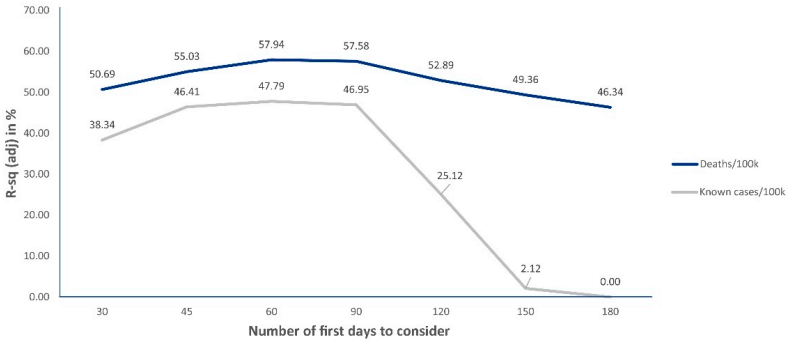

Figure 2. Preliminary linear model fitting results to determine the best time-lapse to address in regression analyses. The first 60 days performed better both for cases per 100 k hab and deaths per 100 k hab.

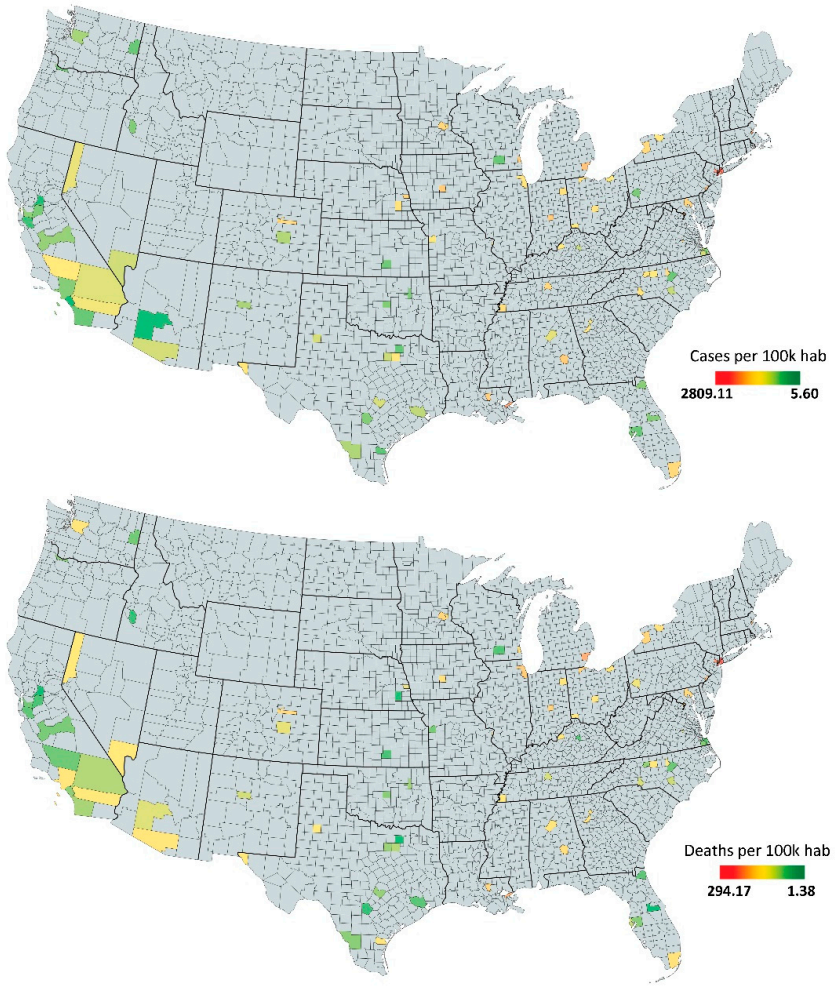

Figure 3. The 93 sampling counties of this study with a color map indicating the number of cases per 100 k inhabitants 60 days after the first case in each county (**above**) and the number of deaths per 100 k inhabitants 60 days after the first death in each county (**below**).

3.2. Best Subsets Regression

The best subsets regression (BSR) is, together with forward stepwise and the lasso, the most popular methods for selecting and estimating parameters in a linear model. While the first two are understood as classical methods in statistics, the lasso is relatively more recent [43]. In a recent work, Hastie, Tibshirani, and Tibshirani (2020) [44] extensively addressed the potentialities and drawbacks of each of these methods, concluding that (1) neither BSR nor the lasso uniformly dominates the other; and (2) for a large proportion of the settings they considered, best subset and forward stepwise perform similarly, with BSR performing better in some specific situations. In turn, Bertsimas, King, and Mazumder (2016) [45] presented empirical comparisons of BSR with other popular variable selection procedures, including the lasso and forward stepwise selection. Their simulations suggested that BSR consistently outperformed the other methods in terms of prediction accuracy.

Thus, in this work, we adopt the (BSR) to test all possible combinations of the independent variables and select the best model according to goodness-of-fit criteria [46–48]. Since we are approaching a social and behavioral data analysis using multivariate regression, we are also interested in understanding the role of our potential predictors (independent variables) in the dynamics of disease spread. Thus, we adopted a correlation analysis, followed by a best subsets regression to determine which of our candidate independent variables (walk score, population density, population size, and the number of days in stay-at-home order) should be considered in our final regression model. This procedure was performed to build two regression models for comparison: one considering the number of cases per 100 k habitants 60 days after the first case in each county and the other considering the number of deaths per 100 k habitants 60 days after the first death in each county. We also performed an analysis considering population size data in its log-transformed state, but the model presented in this paper had a greater r. The goal was to use best subset regressions in the number of known cases per 100 k hab and deaths per 100 k hab against our set of independent variables to determine the most significant dependent and independent variables. Figure 4 illustrates our workflow for selecting the best regression model using the best subsets regression. Thus, we are approaching two conflicting considerations: minimizing the number of predictors to achieve a less expensive model and maximizing the model's explanatory power. In addition to the regressions, we made a multivariable comparison between the counties with the higher and lower number of confirmed cases and deaths per 100 k habitants. These analyses generated preliminary findings that address the following questions: Which urban features matter most? Which can we ignore? How do urban features interact with each other?

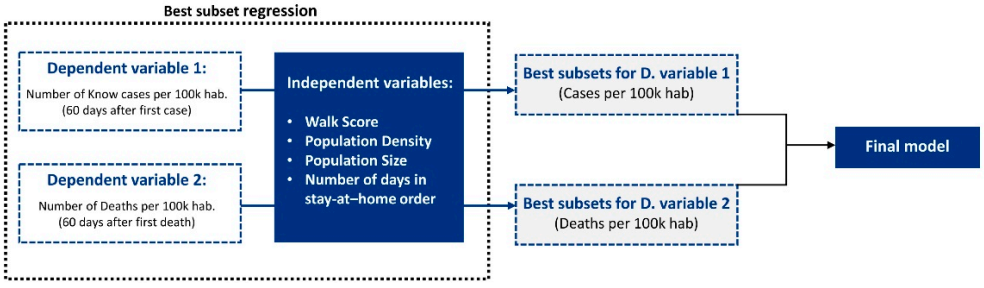

Figure 4. The workflow for building our final regression model: best subset regressions on the number of known cases per 100 k hab and deaths per 100 k hab were used against our set of independent variables to determine the most significant dependent and independent variables.

4. Results

4.1. Correlation Analysis

In order to quantify the degree to which our independent variables are related and avoid biasing the models, we performed a correlation analysis considering all the addressed

independent variables before running the best subsets regressions (Figure 5). Although walk score and population density present a correlation coefficient (r) of 0.582, indicating a moderate positive relationship (when $0.3 < r < 0.7$), there is no strong relationship (when $r \geq 0.7$) between any of the predictors. These correlation thresholds that supported our interpretation were applied together with graphical analysis and are broadly accepted guidelines for interpreting the correlation coefficient [49,50].

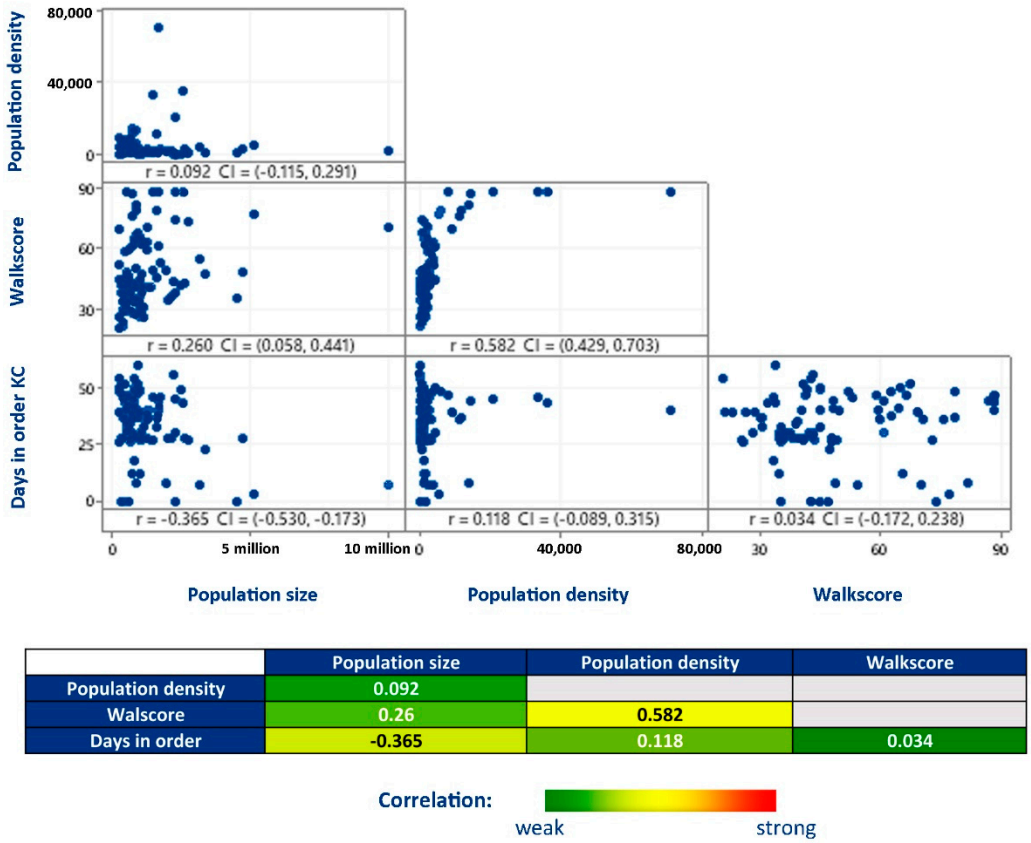

Figure 5. A correlation analysis considering all the addressed independent variables. Although walk score and population density present a moderate positive relationship (0.582), there is no strong relationship ($\geq$0.7) between any of the predictors.

4.2. Best Subsets Regression

Following the check for correlation, our best subset regression analyses results (Tables 1 and 2) show that all models that address the number of deaths per 100 k habitants presented higher values of adjusted R-sq, meaning that they fit the observed data better than the models that address the number of known cases per 100 k habitants. This analysis shows the two best models considering one independent variable, two independent variables, three independent variables, and the model containing all four independent variables. These models were compared to define which independent variables would be addressed in our final model.

When considering the first BSR alone (known cases per 100 k hab), the model addressing all four independent variables provided the highest R-Sq values and the lowest standard errors (S). On the other hand, when considering the second BSR alone (deaths per 100 k hab), the model addressing population density (PD), walk score (W.S.), and the number of days in stay-at-home order (D.O.) as independent variables provided the same R-Sq (adj) as the all-variables model, but with a lower value of standard error (S). Thus, when considering all subsets' possibilities, we chose to build our final regression model considering the number of deaths per 100 k hab (60 days after the first day in each county) versus population density, walk score, and the number of days in stay-at-home order,

since it provided us with the best R-Sq, S, and Mallows' Cp values while addressing a smaller number of independent variables. R-sq informs the fitness of a model, S informs the standard error, and small Mallows' Cp values indicate that the model has slight variance in estimating the accurate regression coefficients and predicting future responses.

Table 1. Best subset regression results considering the number of known cases per 100 k hab as the response. The model with all four independent variables (highlighted) provided the highest R-Sq (adj) and the lowest standard error (S).

Best Subset Regression Results 1—Response Is Know Cases per 100 k hab (after 60 Days from the First Case)					
Vars	**R-Sq**	**R-Sq (adj)**	**R-Sq (pred)**	**Mallows Cp**	**S**
1	39.2	38.5	33.7	29.1	462.24
1	34.8	34.1	0.0	37.4	478.43
2	46.9	45.7	41.0	16.1	434.22
2	46.9	45.7	10.9	16.2	434.36
3	53.0	51.4	20.1	6.5	411.03
3	51.7	50.0	16.9	9.0	416.65
4	54.8	52.7	21.8	5.0	405.32
Vars	**PD**	**WS**	**DO**	**PS**	
1		X			
1	X				
2		X	X		
2	X	X			
3	X	X	X		
3	X	X		X	
4	X	X	X	X	

Table 2. Best subset regression results considering the number of deaths per 100 k hab as the response. The model addressing population density (P.D.), walk score (W.S.), and the number of days in a stay-at-home order (D.O.) as independent variables (highlighted) provided the overall highest R-Sq (adj) and the lowest standard error (S).

Best Subset Regression Results 2—Response Is Deaths per 100 k hab (after 60 Days from the First Death)					
Vars	**R-Sq**	**R-Sq (adj)**	**R-Sq (pred)**	**Mallows Cp**	**S**
1	50.2	49.6	0.0	39.6	42.007
1	49.4	48.9	45.0	41.5	42.309
2	62.9	62.1	24.8	8.9	36.421
2	53.8	52.7	48.9	32.4	40.690
3	65.7	64.5	29.6	3.9	35.261
3	64.4	63.2	26.9	7.3	35.919
4	66.0	64.5	29.8	5.0	35.272
Vars	**PD**	**WS**	**DO**	**PS**	
1	X				
1		X			
2	X	X			
2		X	X		
3	X	X	X		
3	X	X		X	
4	X	X	X	X	

4.3. Final Regression Model

Our analysis shows noteworthy correlations between walkability, population density, and the number of days at stay-at-home order with the number of deaths per 100 k hab, 60 days after the first case in each county (Tables 3 and 4, and Figure 6). We came to the following findings after a normality test and a Box-Cox transformation of $\lambda = 0.5$ to our data. Our regression model provided an R-sq (adj) of 64.85% and a standard error (S) of 2.13467, which can be seen as very significant, especially if we consider that a set of non-measurable social behavior-related features such as how different groups choose to mask, stay home, and take other preventive measures also influence COVID-19 spread. The population density and walk score predictors presented p-values < 0.01, indicating solid evidence of statistical significance, while the number of stay-at-home days predictor presented a p-value < 0.05, indicating moderate evidence of statistical significance [51,52]. Overall, our Pareto chart of the standardized effects shows that walk score's effect, population density's effect, and days in order's effect are more significant than the reference value for this model (1.987), meaning that these factors are statistically significant at the 0.05 level with the current model terms. Following these findings, our residual plot analyses (probability, fits, histogram, and order) validated the model.

Thus, our regression analyses positively correlated deaths per 100 k habitants and all independent variables. It means that as walk score, population density, and the number of days in stay-at-home order increases, these COVID-19 related numbers tend to be higher. Figure 7 depicts the evolution of cases and deaths per 100 k habitants through time, relating these numbers to each predictor and comparing the models for the number of cases and the number of deaths. Although it might seem controversial that the number of deaths increased with the number of days at home, our time-lapse sample, which intentionally addressed the initial stages of the spread, makes it reasonable to assume that places with higher disease spread adopted more robust measures as a reaction. Containment measures have a timing aspect that influences their performance. According to [53], the benefits of a lockdown are seen around 15–20 days before the peak of the epidemic, providing a limited window for public health decision-makers to mobilize and take full advantage of lockdown as an NPI.

Table 3. Final model summary for transformed response (Box-Cox transformation $\lambda = 0.5$).

Regression Equation						
Deaths per 100 k hab^0.5= −2.672 + 0.000130 Population density + 0.1098 Walkscore + 0.0401 Days in order KC						
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2.13467	66.01%	64.85%	631.932	46.44%	407.22	419.13

Table 4. Coefficients for the transformed response.

Term	Coef	S.E. Coef	95% CI	T-Value	*p*-Value
Constant	−2.672	0.918	(−4.496, −0.848)	−2.91	0.005
Population density	0.000130	0.000030	(0.000071, 0.000190)	4.33	0.000
Walkscore	0.1098	0.0155	(0.0791, 0.1406)	7.10	0.000
Days in order KC	0.0401	0.0160	(0.0084, 0.0718)	2.51	0.014

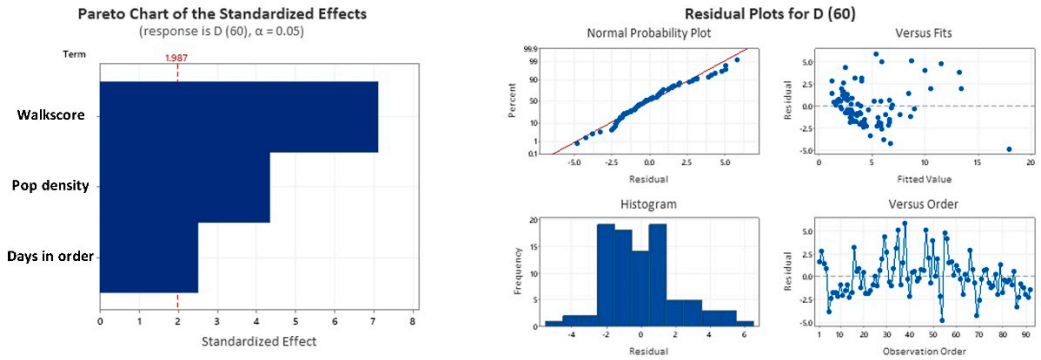

Figure 6. The Pareto chart of the standardized effects depicting the statistical significance of the addresses terms (**left**) and the residual plots for validating the model (**right**).

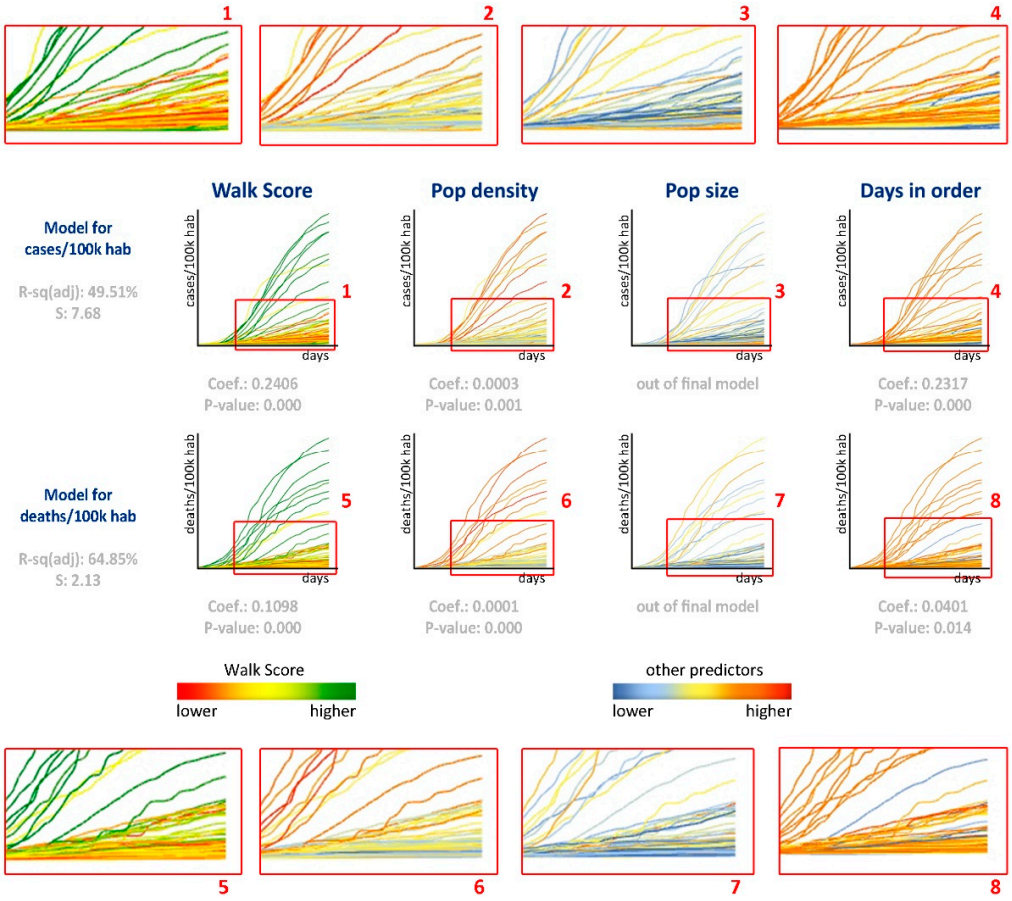

Figure 7. Cases per 100 k hab (**above**) and Deaths per 100 k hab (**below**) evolution in the 60 days after the first case (**above**) and death (**below**). Each line represents one of the analyzed counties. Different predictors weigh the data visualization.

4.4. Discussion

The COVID-19 pandemic and all of the complex data that it generates rely on a simple relationship: contact leads to infection. In this sense, cities are the stage on which contact between people and, therefore, the infection takes place. This preliminary study's findings confirm our hypothesis that certain urban features (population density and walkability) correlate more with COVID-19 spread in the first days of the pandemic than other variables such as overall population size. Despite addressing an initial and limited set of predictor variables, we have found some important correlations (not a causal relationship, but an association to be further explored nonetheless). Considering our research scope, goals, and

hypothesis (the impact of urban features on the disease spread), it is essential to highlight the importance of addressing the early stages of contagion to observe the trends before containment measures had a more significant influence.

Our results suggest a clear positive correlation between Walk Score and the number of deaths/100 k habitants, but it does not mean that the act of walking itself promotes higher contagion rates. Instead, it reinforces that places most likely to congregate amenities and promote encounters are potentially more contagious and require more effective containment actions.

5. Final Remarks: Limitations and Further Developments

5.1. Limitations of This Work

Despite achieving meaningful findings, the authors recognize that this study has some limitations. Disease spread during pandemics can be influenced by a host of social, economic, and behavioral factors in complex ways. In order to achieve more extensive results and increase our model's fitness, it would be essential to address a broader sample of predictor variables, urban features, and urban scales, in addition to a more extensive time-lapse, weighted by other containment measures. Analyses at varying scales could, for example, address disparities within a single city across neighborhoods that are often starkly different despite their geographic proximity. Moreover, variables that capture heterogeneity across and within urban counties are both important. We also acknowledge that variables related to issues such as socioeconomic vulnerability, disproportionate spread in rural areas, where population sizes, population densities, and walkability indicators are small, population density expressed as the proportional population within a set of population density bands, and timing since the first case in the U.S. seem to be important and will be addressed in further developments.

On the other hand, this experiment provided an important basis for future work regarding the impact of urban features on COVID-19 (and similar contagious infections) spreading. Although some works have approached possible correlations of COVID-19 spread, population size, and population density in an American context [30,35], our research is, to the best of our knowledge, among the first to consider walkability as an important urban feature in this regard, since it is correlated with the interaction of people in urban environments as demonstrated in previous studies [20–22]. Our method assumes that the initial spread of the disease is less impacted by containment measures.

5.2. Future Work

In future stages of this research, we plan to address: (i) more urban features (e.g., mixed-use and floor area indexes, network density, volumetric compactness, containment measures, and so on); (ii) more urban scales (cities, zip codes, neighborhoods, and rural areas); (iii) a larger sample of time and cases (the timeline for the first 365 days in United States, for instance), (iv) socioeconomic, ethnic and racial indicators, (v) health-related indicators such as BMI, to identify physical activity levels in different places; (vi) urban mobility indicators; (vii) national and international connection measures and; (viii) data mining and machine learning techniques to retrieve, analyze, and model urban and infection data in different contexts. The expectation is that understanding how these features lead to different modes of social interaction and, consequently, to different dissemination patterns of COVID-19 will help identify appropriate strategies to contain and mitigate the infection and alternative healthcare policies.

More complex sets of features may also require the use of additional tools, such as data mining and machine learning techniques to retrieve, analyze, and model urban and infection data in different contexts. For example, the creation and analysis of an artificial neural network (ANN) might better capture the relationship between urban predictors and quantitative descriptors of COVID-19 spread. ANN is particularly useful when utilizing conventional engineering, or statistical approaches is not possible due to insufficient domain knowledge or the time and resource required makes it impractical.

Several researchers consider walkability and density key aspects of sustainable urbanism since they increase the potential of socioeconomic interactions and optimize energy performance, reducing pollution and resource consumption. However, our findings indicate that these urban features may directly correlate with a greater spread of COVID-19 in urban areas (at least in the United States scenario). In this context, urban planning may face some challenges in post-pandemic times to find answers to the following questions. How to balance the social, environmental, and economic need for such urban features with the need to provide more resilient and healthier cities? If a city can be conceived as a network, how can we ensure that it is flexible, efficient, and yet responsive to health? Which urban features have the greater impact on the spread of infectious diseases, and how can we manage them in this context? We believe that this work is one step in this direction.

Author Contributions: Methodology, F.T.L., N.C.B. and J.P.D.; analysis, F.T.L., N.C.B. and J.P.D.; investigation, F.T.L.; supervision, N.C.B. and J.P.D.; writing—original draft preparation, F.T.L.; writing—review and editing, N.C.B. and J.P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financed in part by the Stuckeman Center for Design Computing and the Stuckeman School of Architecture and Landscape Architecture, The Pennsylvania State University, United States, and by the Brazilian Coordination of Superior Level Staff Improvement - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Andersen, M. *Early Evidence on Social Distancing in Response to COVID-19 in the United States*; Social Science Research Network: Rochester, NY, USA, 2020.
2. U.S. Department of Health and Human Services COVID-19 and Your Health. Available online: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html (accessed on 7 April 2021).
3. Greenstone, M.; Nigam, V. *Does Social Distancing Matter?* Social Science Research Network: Rochester, NY, USA, 2020.
4. Lewnard, J.; Lo, N. Scientific and Ethical Basis for Social-Distancing Interventions against COVID-19—The Lancet Infectious Diseases. *Lancet* **2020**, *20*, 631–633. [CrossRef]
5. Alexander, C. A City Is Not a Tree. *Archit. Forum* **1965**, *122*, 58–62.
6. Hillier, B.; Hanson, J. *The Social Logic of Space*; Cambridge University Press: Cambridge, UK, 1984.
7. Pflieger, G.; Rozenblat, C. Introduction. Urban Networks and Network Theory: The City as the Connector of Multiple Networks. *Urban Stud.* **2010**, *47*, 2723–2735. [CrossRef]
8. Netto, V.M.; Brigatti, E.; Meirelles, J.; Ribeiro, F.L.; Pace, B.; Cacholas, C.; Sanches, P. Cities, from Information to Interaction. *Entropy* **2018**, *20*, 834. [CrossRef] [PubMed]
9. Batty, M. The Coronavirus Crisis: What Will the Post-Pandemic City Look Like? *Environ. Plan. B Urban Anal. City Sci.* **2020**, *47*, 547–552. [CrossRef]
10. United Nations COVID-19 in an Urban World. Available online: https://www.un.org/en/coronavirus/covid-19-urban-world (accessed on 7 April 2021).
11. Farr, D. *Sustainable Urbanism: Urban Design With Nature*; John Willey and Sons: Hoboken, NJ, USA, 2008.
12. Gehl, J. *Cities for People*; Island Press: Washington, DC, USA, 2010.
13. Lima, F.; Paraízo, R.C.; Kós, J.R. Algorithmic Approach towards Transit-Oriented Development Neighborhoods: (Para)Metric Tools for Evaluating and Proposing Rapid Transit-Based Districts. *Int. J. Archit. Comput.* **2016**, *14*, 131–146. [CrossRef]
14. Lima, F.; Montenegro, N.; Paraizo, R.; Kós, J. Urbanmetrics: An Algorithmic-(Para)Metric Methodology for Analysis and Optimization of Urban Configurations. In *Planning Support Science for Smarter Urban Futures*; Lecture Notes in Geoinformation and Cartography; Geertman, S., Allan, A., Pettit, C., Stillwell, J., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 47–64. ISBN 978-3-319-57819-4.
15. Savini, L.; Candeloro, L.; Calistri, P.; Conte, A. A Municipality-Based Approach Using Commuting Census Data to Characterize the Vulnerability to Influenza-Like Epidemic: The COVID-19 Application in Italy. *Microorganisms* **2020**, *8*, 911. [CrossRef] [PubMed]

16. Frank, L.D.; Schmid, T.L.; Sallis, J.F.; Chapman, J.; Saelens, B.E. Linking Objectively Measured Physical Activity with Objectively Measured Urban Form: Findings from SMARTRAQ. *Am. J. Prev. Med.* **2005**, *28*, 117–125. [CrossRef] [PubMed]
17. Buck, C.; Pohlabeln, H.; Huybrechts, I.; De Bourdeaudhuij, I.; Pitsiladis, Y.; Reisch, L.; Pigeot, I. Development and Application of a Moveability Index to Quantify Possibilities for Physical Activity in the Built Environment of Children. *Health Place* **2011**, *17*, 1191–1201. [CrossRef]
18. Dobesova, Z.; Krivka, T. Walkability Index in the Urban Planning: A Case Study in Olomouc City. In *Advances in Spatial Planning*; Burian, J., Ed.; InTech: Rijeka, Croatia, 2012; pp. 179–196.
19. Lima, F. Urban metrics: (para)metric system for analysis and optimization of urban configurations. Ph.D. Thesis, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2017.
20. Brewster, M.; Hurtado, D.; Olson, S.; Yen, J. Walkscore Com: A New Methodology to Explore Associations between Neighborhood Resources, Race, and Health. In Proceedings of the 137th APHA Annual Meeting and Exposition 2009, Philadelphia, PA, USA, 7–11 November 2009.
21. Walkscore Methodology. Available online: https://www.walkscore.com/methodology.shtml. (accessed on 16 June 2016).
22. Carr, L.; Dunsigen, I.; Marcus, B. Validation of Walk Score for Estimating Access to Walkable Amenities. *Br. J. Sports Med.* **2011**, *45*, 1144–1158. [CrossRef]
23. United Nations 68% of the World Population Projected to Live in Urban Areas by 2050, Says UN | UN DESA | United Nations Department of Economic and Social Affairs. Available online: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html (accessed on 7 April 2021).
24. Tarwater, P.M.; Martin, C.F. Effects of Population Density on the Spread of Disease. *Complexity* **2001**, *6*, 29–36. [CrossRef]
25. Hu, H.; Nigmatulina, K.; Eckhoff, P. The Scaling of Contact Rates with Population Density for the Infectious Disease Models. *Math. Biosci.* **2013**, *244*, 125–134. [CrossRef]
26. Kraemer, M.U.G.; Perkins, T.A.; Cummings, D.A.T.; Zakar, R.; Hay, S.I.; Smith, D.L.; Reiner, R.C. Big City, Small World: Density, Contact Rates, and Transmission of Dengue across Pakistan. *J. R. Soc. Interface* **2015**, *12*, 20150468. [CrossRef]
27. Dantzig, G.; Saaty, T. *Compact City: A Plan for a Liveable Urban Environment*; W. H. Freeman: San Francisco, CA, USA, 1973.
28. Rogers, R. *Cities for a Small Planet*; Westview Press: Boulder, CO, USA, 1997.
29. Chakrabarti, V. *A Country of Cities*; Metropolis Books: New York, NY, USA, 2013.
30. Schläpfer, M.; Bettencourt, L.M.A.; Grauwin, S.; Raschke, M.; Claxton, R.; Smoreda, Z.; West, G.B.; Ratti, C. The Scaling of Human Interactions with City Size. *J. R. Soc. Interface* **2014**, *11*, 20130789. [CrossRef]
31. Stier, A.; Berman, M.G.; Bettencourt, L. *COVID-19 Attack Rate Increases with City Size*; Social Science Research Network: Rochester, NY, USA, 2020.
32. Liu, L. Emerging Study on the Transmission of the Novel Coronavirus (COVID-19) from Urban Perspective: Evidence from China. *Cities* **2020**, *103*, 102759. [CrossRef]
33. Peng, Z.; Wang, R.; Liu, L.; Wu, H. Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 402. [CrossRef]
34. Xie, J.; Zhu, Y. Association between Ambient Temperature and COVID-19 Infection in 122 Cities from China. *Sci. Total Environ.* **2020**, *724*, 138201. [CrossRef] [PubMed]
35. Carozzi, F. *Urban Density and COVID-19*; Social Science Research Network: Rochester, NY, USA, 2020.
36. Oishi, S.; Cha, Y.; Schimmack, U. The Social Ecology of COVID-19 Cases and Deaths in New York City: The Role of Walkability, Wealth, and Race. *Soc. Psychol. Personal. Sci.* **2021**, *12*, 1457–1466. [CrossRef]
37. Dasgupta, S.; Bowen, V.B.; Leidner, A.; Fletcher, K.; Musial, T.; Rose, C.; Cha, A.; Kang, G.; Dirlikov, E.; Pevzner, E.; et al. Association Between Social Vulnerability and a County's Risk for Becoming a COVID-19 Hotspot—United States, June 1–July 25, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **2020**, *69*, 1535–1541. [CrossRef] [PubMed]
38. Rocha, R.; Atun, R.; Massuda, A.; Rache, B.; Spinola, P.; Nunes, L.; Lago, M.; Castro, M.C. Effect of Socioeconomic Inequalities and Vulnerabilities on Health-System Preparedness and Response to COVID-19 in Brazil: A Comprehensive Analysis. *Lancet Glob. Health* **2021**, *9*, 782–792. [CrossRef]
39. Subbaraman, N. Why Daily Death Tolls Have Become Unusually Important in Understanding the Coronavirus Pandemic. Available online: https://www.nature.com/articles/d41586-020-01008-1 (accessed on 31 March 2021).
40. Kunz, J.; Propper, C. *"Does Higher Hospital Quality Save Lives? The Association between" "COVID-19 Deaths and Hospital Quality in the USA*; Social Science Research Network: Rochester, NY, USA, 2020.
41. USAFacts US COVID-19 Cases and Deaths by State | USAFacts. Available online: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (accessed on 7 April 2021).
42. NBC news Report: Here Are the Stay-at-Home Orders in Every State. Available online: https://www.nbcnews.com/health/health-news/here-are-stay-home-orders-across-country-n1168736 (accessed on 2 April 2021).
43. U.S. Census Census.Gov. Available online: https://www.census.gov/ (accessed on 7 April 2021).
44. Hastie, T.; Tibshirani, R.; Tibshirani, R. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Stat. Sci.* **2020**, *35*, 579–592. [CrossRef]
45. Bertsimas, D.; King, A.; Mazumder, R. Best Subset Selection via a Modern Optimization Lens. *Ann. Stat.* **2016**, *44*, 813–852. [CrossRef]

46. Hofmann, M.; Gatu, C.; Kontoghiorghes, E.J. Efficient Algorithms for Computing the Best Subset Regression Models for Large-Scale Problems. *Comput. Stat. Data Anal.* **2007**, *52*, 16–29. [CrossRef]
47. Zhang, Z. Variable Selection with Stepwise and Best Subset Approaches. *Ann. Transl. Med.* **2016**, *4*, 136. [CrossRef]
48. Kwong, Y.D.; Mehta, K.M.; Miaskowski, C.; Zhuo, H.; Yee, K.; Jauregui, A.; Ke, S.; Deiss, T.; Abbott, J.; Kangelaris, K.N.; et al. Using Best Subset Regression to Identify Clinical Characteristics and Biomarkers Associated with Sepsis-Associated Acute Kidney Injury. *Am. J. Physiol.-Ren. Physiol.* **2020**, *319*, F979–F987. [CrossRef]
49. Ratner, B. The Correlation Coefficient: Its Values Range between +1/−1, or Do They? *J. Target. Meas. Anal. Mark.* **2009**, *17*, 139–142. [CrossRef]
50. Dancey, C.; Reidy, J. *Statistics Without Maths for Psychology*, 7th ed.; Pearson: London, UK, 2017; ISBN 978-1-292-12888-7.
51. Burdette, W.J. *Planning and Analysis of Clinical Studies*; Thomas: New York, NY, USA, 1970.
52. Arsham, H. Kuiper's P-Value as a Measuring Tool and Decision Procedure for the Goodness-of-Fit Test. *J. Appl. Stat.* **1988**, *15*, 131–135. [CrossRef]
53. Oraby, T.; Tyshenko, M.G.; Maldonado, J.C.; Vatcheva, K.; Elsaadany, S.; Alali, W.Q.; Longenecker, J.C.; Al-Zoughool, M. Modeling the Effect of Lockdown Timing as a COVID-19 Control Measure in Countries with Differing Social Contacts. *Sci. Rep.* **2021**, *11*, 3354. [CrossRef]