Variation in Covid-19 Cases Across New York City

May 20, 2020

Awi Federgruen, Sherin Naha

Abstract

The number of confirmed COVID -19 cases, relative to the population size, has varied greatly throughout the United States and, even, within the same city. In different zip codes in New York City, the epicentre of the epidemic, the number of cases per 100,000 residents has varied between 437 and 4227, i.e., in a 1:10 ratio. To guide policy decisions regarding containment and reopening of the economy, schools and other institutions, it is vital to identify which factors drive this large variation.

This paper reports on a statistical study of the incidence variation across New York City, conducted with data at zip code granularity. Among many socio-economical and demographic measures considered, the average household size emerges as the single most important explanatory variable: an increase of the *average* household size by one member accounts, in our final model specification, for at least 876 cases, a full 23% of the span of incidence numbers, at a 95% confidence level.

The percentage of the population above the age of 65, as well as that below the poverty line, are additional indicators with a significant impact on the case incidence rate, along with their interaction term.

Contrary to common belief, population density, per se, fails to have a significantly positive impact. Indeed, population densities and case incidence rates are *negatively* correlated, with a -33% correlation coefficient.

Our model specification is anchored on a basic and established epidemiological model that explains the importance of household sizes on R0, the basic reproductive number of an epidemic.

Our findings support implemented and proposed policies to quarantine pre-acute and post-acute patients, , as well as nursing home admission policies.

I. Introduction

The world continues to search for a fundamental understanding of the dynamics of the current pandemic. For example, we try to understand why, in the USA, as of May 17, 192000 confirmed cases have arisen in New York City alone, while the country-wide total has been mercifully restricted to 1.5 million, a staggering number, nevertheless. In other words, to date, a city representing 2.5% of the nationwide population experienced 12.8 % of the confirmed reported cases. Identifying the main drivers behind the contagion process has important implications for the public policies we should implement to contain the current epidemic and mitigate the generally expected "second wave".

It has been widely conjectured that *population density* is a main driving force in the contagion process. This theory has some, a priori, intuitive appeal. After all, the number of infections in a given region depends on the basic reproductive number R0, defined as the average number of cases directly generated by a single case, in a population in which all individuals are susceptible. This reproductive number, in turn, depends, *in part*, on the number of distinct individuals, a single case has physical contact with, during the time interval in which he or she is contagious. Ceteris paribus, the latter can be expected to be positively correlated with the population density.

However, the theory is challenged, first of all, on an international basis. Many cities with far greater or comparable population density than New York City's 10198 residents per square kilometre, have experienced much lower absolute and relative case rates. These include , for example, Manilla (46.128 residents/sq. km), Baghdad (32,874 residents/sq. km), Mumbai (32,303 residents/sq. km), Seoul (16,000 residents/sq. km), Mexico City (9,800 residents/sq. km), and Singapore (8,358 residents/sq. km).

Such reverse patterns may, possibly, be explained by, ex ante differences in international traffic patterns in and out of the country, affecting the cluster of "imported" cases, or the specific containment and testing policies adopted by the respective governments. However, similar reverse patterns arise when comparing states within the USA as well: California faces one of the lowest COVID mortality rates (8 per 100,000 residents), but has one of the highest population densities (251.3 residents per square mile); in contrast, the state of Louisiana has 49 fatalities per 100,000 residents, while its population density is, approximately, 2.5 times lower than that of California.

And what should be made of stark differences within New York City, itself? Wadhera et al. (JAMA, April 29, 2020) reported , recently, that among all five boroughs in the city, Manhattan had by far the *lowest* number of hospitalizations per 100,000 residents, all while facing the *largest* population density, 2.5 times the citywide average (25106 vs 10198 residents/sq. km). (The percentage of confirmed cases that requires hospitalization is remarkably robust, across the country, so the same relationships pertain with respect to the incidence rate of confirmed cases). See Figure 1 below. Last but not least, data at zip code granularity, show that the rates of confirmed cases, and the population densities are significantly *negatively* correlated, with a correlation coefficient of – 33%, see Table 2, below.

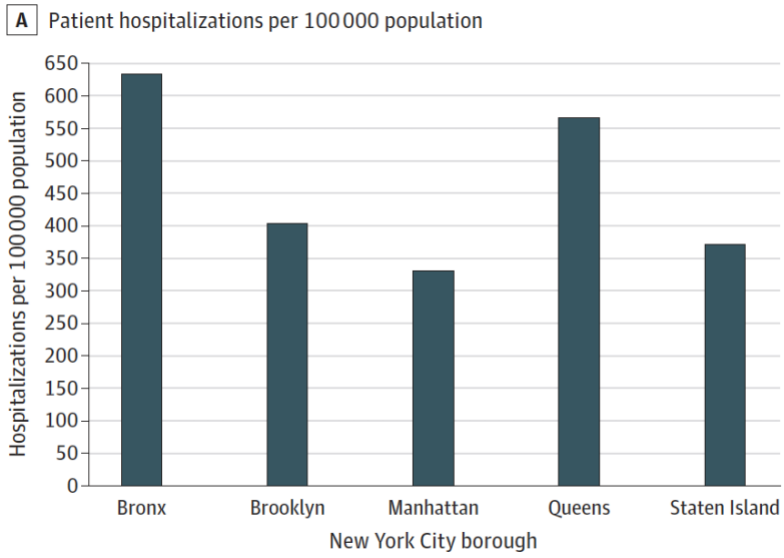

Fig1. **Hospitalization rates due to Coronavirus Disease 2019 (COVID 19) , from Wadhera et al. (JAMA, April 29, 2020)**

Several authors have, indeed, cautioned against the general emphasis on *population density* as a prime explanatory variable. These include Barr and Tassier (Scientific American, April 17, 2020) and Bassett (New York Times, May 15 2020).

First and foremost, it is understood that populations consist of "households" and that the inter-household and intra-household dynamics of an epidemic differ fundamentally. In particular, the contact rate between a pair of individuals tends to be much greater when the individuals share the same house-hold than is the case for individuals from different households. Starting with the seminal papers by Bartoszynski (1972) and Becker (1977), this observation has been at the core of many epidemiological models. See Section 2 for a brief description of a seminal model by Becker and Dietz (1995).

The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team in China (China CDC Weekly, 2020) reported that 80% of transmission clusters of the coronavirus occurred *within households*. Nearly three weeks into the initial coronavirus outbreak in Wuhan, China, the government instituted social distancing and travel bans. This had the substantial impact of reducing the reproductive number R0 from 3.88 to 1.26. However, it wasn't until Wuhan instituted a centralized quarantine where anyone with COVID-like symptoms (e.g. cough, fever, etc.) would be centrally quarantined at hotels or dormitories, that the estimate for R0 was reduced to 0.32, a 75% improvement over social distancing alone! Indeed, this intervention was the final one in achieving full containment of the virus in China.

And Manhattan, while the most densely populated of all New York City boroughs, is also the one with the *smallest average household size.* This demographic indicator shows large variation across zip codes, from a minimum of 1.57 to a maximum of 3.97, see Table 1. It is also highly *positively* correlated with the number of confirmed cases per 100,000 residents, see Table 2. Thus, any

statistical model explaining the variation in case intensities across the City, should include the average household size as an explanatory variable.

More broadly, we need to identify demographic and socio-economic indicators that have significant predictive power. New York City is an ideal arena to pursue this investigation; as mentioned, the city has, by far, the highest confirmed incidence rate in the country, but it also exhibits remarkable diversity with respect to many of the potentially relevant demographic and socio-economic factors. Combining various data sources, described below, we have been able to assess these factors, on a zip-code by zip-code basis. (We have identified 174 zip-codes for which data were fully available.) See Table 1 for a few examples.

	Minimum	Maximum	Average
# confirmed cases per 100000 residents	436.91	4226.51	2077.79
Population density= # residents per sq. mile	1389.08	126067.69	38480.75
Average household size	1.57	3.97	2.65
Percentage below poverty line	2.20%	45.90%	16.83%
Percentage above the age of 65	0.46%	28.98%	14.26%

Table 1: Descriptive statistics by NYC zip-code

II. A basic household based epidemiology model

Becker and Dietz (1995) consider a population with households of varying sizes. Let μ_H, σ_H and $cv_H = \sigma_H / \mu_H$ denote the mean , standard deviation and coefficient of variation of the distribution of household sizes, respectively. Let b denote the mean number of infectious contacts that an individual makes with individuals outside her own household during her entire infectious period. An infectious contact is one that is close and intensive enough to result in disease transmission.

Becker and Dietz (1995, p.211) derive the following closed form expression for R0, the basic reproductive number of the epidemic:

$$R_0 = \frac{b}{2}\left[1 + \left(1 + \frac{4(\mu_H(1 + cv_H^2) - 1)}{b}\right)^{\frac{1}{2}}\right] \quad (1)$$

In a population where all individuals live by themselves, $\mu H = 1$ and $cv_H = 0$, so that $R0 = b$, as in elementary models that do not account for household clusters. In contrast, when people live in households of size 4, for example married couples with 2 children, $\mu_H = 4$ and , if $b >= 1$, i.e., an average individual infects at least one individual outside her own household:

$$R_0 = 0.5\left(b + \sqrt{b^2 + 12b}\right) = b + 0.5\left(\sqrt{b^2 + 12b} - b\right)$$

$$= b + \frac{6b}{(\sqrt{b^2 + 12b} + b)} = b + \frac{6}{\sqrt{1 + 12\,b^{-1}} + 1} \geq b + 1.3$$

, adding at least a full 1.3 points to the R0 value! And when there is variability in the household sizes, as there typically is, the household effect is even greater, see (1). For example, if the household distribution approaches a Geometric distribution,

$cv_H = \sqrt{1 - \frac{1}{\mu_H}} = \sqrt{0.75} = 0.87$, $R_0 = 0.5\left(b + \sqrt{b^2 + 24b}\right) = b + \frac{12}{\sqrt{1+24b^{-1}}+1} \geq b + 2$

Population density, by itself, may affect R0 via the parameter b, depending on the nature of the interactions among individuals in different households. But if it does, (1) shows that there is a strong interaction effect with the average household size (and its coefficient of variation) as well.

III. Methods:

The numbers of confirmed COVID19 cases per zip code were obtained from the New York City Department of Health's GitHub repository. The data are maintained in a CSV (comma separated value) file, (see: https://github.com/nychealth/coronavirus-data/blob/master/tests-by-zcta.csv)

Data pertaining to population sizes and population densities per zip code were retrieved from http://zipatlas.com/us/ny/zip-code-comparison/population-density.htm.

Lastly, all demographic and socio-economic zip-code characteristics were extracted from a CSV database created by Buzzfeed, see https://github.com/BuzzFeedNews/2020-05-covid-city-zip-codes/blob/master/data/processed/census/zip_census_data.csv. The database, in turn, was created from 2018 5-year American Community Service (ACS) estimates that were obtained from the USA Census Bureau. The following is a list of age and economic status factors contained in this data base:

Age:

- age_60_and_over
- age_65_and_over
- age_75_and_over
- median_age

Economic:

- pct_more_than_one_occupant_per_room
- pct_below_poverty_level
- household_median_income
- household_mean_income

We applied standard linear regression to estimate best fit regression equations, assuming independent noise terms, all with an identical Normal distribution. Based on the empirical and theoretical observations above, we chose a model specification with (a) *population density* and (b) *average household size* as potential explanatory variables, along with (c) an *interaction term* between them, (d) *the percentage of the population above the age of 65*, and (e) *the percentage below the poverty line*. The relevance of the latter two demographics is both intuitive and explained in our Discussion section. In the remainder of this paper, we employ the following variable names:

- *population_density:* Population density=# residents per sq. mile
- *avg_household_size:* Average household size
- *pct_below_poverty_line:* Percentage below poverty line
- *pct_age>65:* Percentage above the age of 65
- *cases_per_100k:* # confirmed cases/100000 residents

As mentioned, the correlation between *population density* and *cases_per_100k* is -33%. A negative –albeit weaker-correlation (-12%) continues to prevail between the dependent variable *cases_per_100k* and the standard interaction term *population density* avg_household_size.* We therefore specified the interaction term as *population density**(*avg_household_size)*$_2$which *is* positively correlated with *cases_per_100k.*

	Population density= # residents per sq. mile	Average household size	Percentage below poverty line	Percentage above the age of 65	# confirmed cases/100000 residents
Population density=# residents per sq mile	1	-0.324	0.285	-0.064	**-0.325**
Average household size	-0.324	1	0.189	-0.207	**0.622**
Percentage below poverty line	0.285	0.189	1	-0.367	0.229
Percentage above the age of 65	-0.064	-0.207	-0.367	1	0.092
# confirmed cases/100000 residents	**-0.325**	**0.622**	0.229	0.092	1

Table 2: Correlation Coefficients among the variables

IV. Results:

As suggested by the correlation numbers in Table 2, average household size is, by far, the single most important explanatory variable in explaining the variation in case incidence rates across the

City. Indeed, it is able to explain 62% of the variation among these incidence rates, by itself. When this variable is used as the sole explanatory variable, the following equation provides the best fit.

 # confirmed cases/100,000=450.5+986.9 (average household size -1)*

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	450.51	163.46	2.76	0.006	127.86	773.154542
avg_household_size_adj	986.90	94.79	10.41	5.41E-20	799.80	1174.01

Table 3: Regression results with average household size as the single explanatory variable; avg_household size_adj = average household size-1.

The coefficient in front of the average household size variable is highly significant. On the basis of this equation, one would estimate that a zip code A with one more average household member than an otherwise comparable zip code B would experience an addition in its case rate by 986.9 cases.

However, a significantly superior fit can be obtained by adding the remaining explanatory variables, see Table 4. In this specification, 72% of the variation in case rates is explained (Multiple R=72%), while R_2 and the Adjusted R_2 are 51% and 50% , respectively, resulting in the following regression equation:

 # confirmed cases/100,000 = -284.4 + 896.8 (average household size -1)+ 49.9*percentage above age 65 + 24.4*percentage below poverty level - 0.006*population density – 3.52E-05*(population density*(average household size -1)$_2$)*

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-284.40	353.05	-0.805	0.42	-981.40	412.59
pop_density	-0.006	0.003	-1.81	0.07	-0.01	0.001
avg_household_size_adj	896.83	180.56	4.97	1.66E-06	540.36	1253.29
pct_below_poverty_level	24.37	5.36	4.55	1.03E-05	13.79	34.942533
pct age>65	49.94	9.80	5.10	9.29E-07	30.59	69.2839432
pop_density* (avg_household_size_adj)$_2$	-3.52E-05	0.001	-0.03	0.98	-0.002	0.00250162

Table 4: Full Regression Results without Interaction Effects

The average household size continues to be the main determinant of the number of reported COVID-19 infections. In contrast, neither population density, nor its interaction term with the average household size, have a significant impact on the case intensity. (The intercept value is insignificant as well).

Average household size, the percentage of the population above the age of 65 and the percentage below the poverty line, all have *positive* coefficients that are significant at a very high confidence level. When accounting for the other demographic variables, the coefficient of the average household size variable is slightly reduced to 896.8 but one continues to conclude at a 95% confidence level, that a difference of but one extra average household member, ceteris paribus, amounts to an increase of at least 540.4 cases, per 100,000 residents.

An increase of the percentage of senior residents by one percentage point, augments, ceteris paribus, the case rate by approximately 50 cases, more than double the effect of an increase of the poverty rate by a single percentage point. The significance of both demographic factors is intuitive, as well; senior residents may be equally susceptible to getting infected as other age brackets. However, they are at increased risk to develop significant symptoms and complications, thus disproportionally contributing to the counts of *confirmed cases*, the only counts we have, at this point. (A recent study of the population in Santa Clara county by Bendavid et al. (April 30, 2020) estimates that the number of confirmed cases, in that county, may have been as small as 2% of the total number of infected individuals.) Likewise, the a priori health status of individuals in the lowest income bracket, is, typically, inferior to that among higher income brackets. Infected individuals with incomes below the poverty line, therefore contribute, disproportionally, to the counts of confirmed cases, as well. Additionally, a much larger percentage of this population segment is employed in on location as opposed to remote –at home- jobs, thus significantly more susceptible to infections.

Finally, we have checked for the presence of other interaction effects, omitting the population density related variables which were found to have an insignificant impact. The only significant interaction effect is that between the percentage of the population below the poverty line, and the percentage of senior residents. See Table 5 for the revised regression results. (The addition of the interaction term has negligible impact on the R_2 or adjusted R_2 value).

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-549.53	246.39	-2.23	0.027	-1035.93	-63.14
avg_household_size_adj	1051.69	89.00	11.82	7.17E-24	875.99	1227.38
pct_below_poverty_level	25.48	5.43	4.692	5.57E-06	14.76	36.20
pct age>65	48.62	9.73	4.997	1.44E-06	29.41	67.82
pct<PL*pct_age>65	-0.004	0.001	-3.03	0.003	-0.006	-0.001

Table 5 Full Regression with interaction effect, but population density omitted

Thus, after omitting the population density related variables . hose impact was found to be statistically insignificant, our estimate for the coefficient of avg_household_size_adj has a much smaller standard error, resulting in a much narrower 95% confidence interval. The point estimate of this coefficient is now 1051.7, and above 876, at a 95% confidence level. The revised model specification has minimal impact on the estimated coefficients of the renaming explanatory variables.

V. Discussion:

Average household size emerges as the single most important demographic variable when explaining the large variation in infection rates throughout New York City. Depending on which of the two specifications in Tables 4 and 5 is selected, we estimate that an additional single household member increases the number of cases by 892 or 1051.

This result is quite intuitive. It confirms both the above empirical findings and its theoretical underpinnings, in Section II.

Of course, the same contagion potential exists in other settings where a significant number of individuals reside together in the same dwelling, for example dormitories and nursing homes. Girvan and Roy (FREOPP, May 2020) report that, in the USA, 40% of Covid-19 induced deaths occurred in nursing homes and assisted living facilities. Their residents face the simultaneous hazard of living in close proximity to many others, as well as belonging to an age group with significantly increased potential for significant symptoms and hence confirmation of their infections . Similarly, Comas-Herrera et al. (May 3, 2020) document that in Australia, Belgium, Canada, Denmark, France, Germany, Hong Kong, Hungary, Ireland, Israel, Norway, Portugal, Singapore, and Sweden, 49.4 % of reported COVID-19 fatalities took place in nursing homes and related facilities.

The percentage of the population above the age of 65, indeed, emerges as a second significant explanatory variable in our study; here we estimate that each additional percentage point contributes 50 cases per 100,000. As explained, the significance of the percentage under the poverty line is also intuitive, with each percentage point contributing an estimated 24 cases: the health status of this segment of the population is generally poorer, enhancing infections with significant symptoms; moreover, a far larger percentage in this income bracket is employed as essential workers with limited abilities for social distancing; for both reasons, individuals in this segment contribute more intensively to the case count. The second factor pertains almost exclusively to individuals below retirement age, a possible explanation for the significantly negative coefficient of the interaction term in the final regression equation, see Table 5.

Our results about the large importance of household size, support such policy initiatives as quarantine policies for infected individuals, either immediately upon being identified as infected or post-hospitalization, see, for example, Chan et al. (Business Insider, April 25 2020). Based on their model in Chan et al.(2020b), the authors show " that the time to reopening [of the economy]

can be shortened by 11%. In addition, assuming symptomatic people are infectious, if 50% of them are quarantined before getting sick enough to go to the hospital, we can reduce the risk of developing severe COVID illness and the time till reopening can be shortened by 86%." It also supports the May 11 decision by New York State to cancel its mandate requiring nursing homes to accept COVID 19 patients.

Future research should try to identify data on other crowding factors, such as number of individuals residing in nursing homes or dormitories.

VI. References

Barr, J., T. Tassier. (2020, April 17). *Are Crowded Cities the Reason for the COVID-19 Pandemic?* Retrieved from Scientific American: https://blogs.scientificamerican.com/observations/are-crowded-cities-the-reason-for-the-covid-19-pandemic/

Bartoszynski, R. (1972). On a certain model of an epidemic. *Appl. Math,* 13(2):139-151.

Bassett, M. T. (2020, May 15). *Just Because You Can Afford to Leave the City Doesn't Mean You Should*. Retrieved from NY Times: https://www.nytimes.com/2020/05/15/opinion/sunday/coronavirus-cities-density.html

Becker, N. G. (1977). On a general stochastic epidemic model. *Theor. Popul. Biol*, 11:23-36.

Becker, N. G., K. Dietz (1995). The Effect of Household Distribution on Transmission and Control of Highly Infectious Diseases. *Mathematical Biosciences,* 127:07-219

Bendavid, E. ,B. Mulaney, N. Sood, S. Shah, E. Ling, R. Bromley-Dulafano, C. Lai, Z. Weissberg, R. Saavedra-Walker, J. Tedrow, D. Tversky, A. Bogan, T. Kupeic, D. Eichner, R. Gupta, J. Ioannidis, J. Bhattacharya (2020, April 30). COVID-19 Antibody Seroprevalence in Santa Clara County, California. *MedRxiv* https://doi.org/10.1101/2020.04.14.20062463

Chan, C., J. Dong, K. Xu (2020). Managing COVID19 outside of the hospital walls: The potential benefits of Centralized Quarantine. *Columbia Business School Working Paper*.

Chan, C., J. Dong, K. Xu (2020, April 25). *Using hotel for coronavirus patients would be a massive help in fighting the pandemic.* Retrieved from Business Insider: https://www.businessinsider.com/fight-covid19-pandemic-use-hotels-coronavirus-patients-2020-4

Comas-Herrera, A., J. Zalakain, C.Litwin, A.T. Hsu, N. Lane, J. L. Fernandez (2020, May 3). *Mortality associated with COVID-19 outbreaks in care homes: early international evidence.* Retrieved from LTC Covid: https://ltccovid.org/wp-content/uploads/2020/05/Mortality-associated-with-COVID-3-May-final-6.pdf

Girvan, G. (2020, May 7). *Nursing Homes & Assisted Living Facilities Account for 40% of COVID-19 Deaths*. Retrieved from FREOPP: https://freopp.org/the-covid-19-nursing-home-crisis-by-the-numbers-3a47433c3f70

The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (2020, February). *Vital Surveillances: The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020*. Retrieved from Chine CDC Weekly: http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

Wadhera, R. K., P. Wadhera, P. Gaba, J.F. Figueoa, K. E. J. Maddox,R.W. Yeh, C. Shen (April 29, 2020). Variation in COVID-19 Hospitalizations and Deaths Across NewYork City Boroughs. *JAMA.*

Acknowledgement

We express our gratitude to Dr. Daniel Berman, Infectious Disease Specialist at Montefiore Hospital, Professor Jing Dong of the Columbia Business School and Professor Edward Kaplan of the Yale School of Management and Professor of Public health at Yale for their insightful suggestions and comments with respect to earlier drafts of this paper

Awi Federgruen is the Charles E. Exley Professor of Management at the Graduate School of Business at Columbia University.

Sherin Naha is a Graduate Student in the Management Science & Engineering Masters Program at Columbia University