Using Demographic Pattern Analysis to Predict COVID-19 Fatalities on the US County Level

KLAUS MUELLER and ERIC PAPENHAUSEN, Akai Kaeru LLC, Stony Brook, NY USA

Unlike pandemics in the past, COVID-19 has hit us in the midst of the information age. We have built vast capabilities to collect and store data of any kind that can be analyzed in myriad ways to help us mitigate the impact of this catastrophic disease. Specifically for COVID-19, data analysis can help local governments to plan the allocation of testing kits, testing stations, and primary care units, and it can help them in setting guidelines for residents, such as the need for social distancing, the use of face masks, and when to open local businesses that enable human contact. Further, it can also lead to a better understanding of pandemics in general and so inform policy makers on the regional and national level. All of this can save both cost and lives. In this article, we show the results of an ongoing study we conducted using a prominent regularly updated dataset. We used a pattern mining engine we developed to find specific characteristics of US counties that appear to expose them to higher COVID-19 mortality. Furthermore, we also show that these characteristics can be used to predict future COVID-19 mortality.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **Visualization systems and tools**; **Visual analytics**; • **Computing methodologies** → **Causal reasoning and diagnostics**;

Additional Key Words and Phrases: Pattern analysis, explainable AI, subspace clustering, visual analytics, predictive analysis, machine learning

ACM Reference format:
Klaus Mueller and Eric Papenhausen. 2020. Using Demographic Pattern Analysis to Predict COVID-19 Fatalities on the US County Level. *Digit. Gov.: Res. Pract.* 2, 1, Article 12 (December 2020), 11 pages.
https://doi.org/10.1145/3430196

1 INTRODUCTION

The impact of a massive disaster like a pandemic differs from other more normal causes of death, such as cancer, heart disease, car accidents, and the like. What makes them so different is their sudden appearance that claims many lives in a short span of time. People get infected and die almost at random, which causes great societal disruption. However, lacking concrete precedence, governments are left without proper guidelines on how to lessen and control the disaster's massive human and economic losses.

This research was partially funded by NSF SBIR grant #1926949, the New York State SPIR program, and the Stony Brook University Sensor CAT program.

Author's addresses: K. Mueller and E. Papenhausen, Akai Kaeru LLC, 1500 Stony Brook Road, Stony Brook, NY 11794-6040, USA; emails: {mueller, epapenha}@akaikaeru.com.

2639-0175/2020/12-ART12
https://doi.org/10.1145/3430196

This is where real-time data analytics can come to the rescue. Thankfully, and in contrast to previous such disastrous events, there is no shortage of data, both on the underlying demographics as a whole and on the spread of the disease in particular. In this article,we report on a study we conducted using a state-of-the-art machine learning approach we developed recently. Our method can identify the specific characteristics of US counties that appear to expose them to higher COVID-19 mortality, and then utilize these profiles to predict their prospects of future COVID-19 mortality as the disease spreads.

2 RELATED WORK

COVID-19 has invigorated scientific research [1]. At the time of writing this article the Allen Institute for AI has compiled a database of over 130,000 research articles on the novel coronavirus [2] and the popular SSRN COVID-19 paper repository held over 5,000 pre-prints [3] Apart from clinical themes, the prediction of infections, fatalities, required resources, and so on, have been extensively studied. Many approaches (such as References [4, 5, 6]) are based on the mechanistic Susceptible–Exposed–Infectious–Recovered (SEIR) compartment simulation model that, at the process level, mimics the way COVID-19 spreads. Popular are also statistical forecasting models such as that by U Washington's Institute for Health Metric and Education (IHME) [7]. It uses a mixed effects non-linear regression model to fit a curve to data from world-wide geographical locations to create projections of infections, death rates, and health resource demands at the local level. The IHME model has become quite popular, in part due to its simple yet effective interactive visual dashboard tools [10]. A smaller number of papers propose the use of AI-inspired models (see References [8, 9]).

2.1 Simulation vs. Statistical Models

Mechanistic models are attractive, since they allow one to simulate the effects of different mitigation measures, such as quarantining, social distancing, school closings, and so on. However, the model's many parameters require accurate estimates of the population in each compartment and their transition rates. This is challenging, since there still is much uncertainty on the mechanics by which the virus spreads, how many people are in each compartment at a given time, and how people interact with each other socially. An important parameter is the disease reproduction rate, R_0, which gauges the number of people infected per case. It is rather sensitive to small changes; a 10% increase in R_0, say, from 2.3 to 2.5, can result in an 80% increase in the predicted infections [10]. Hence, these models require constant monitoring and parameter tweaking to enable long-term predictions of the impact of virus mitigation and re-opening strategies [11].

Statistical methods, such as IHME, do not require tedious parameter tweaking, but since they do not model how transmission occurs they are not well suited for long-term predictions beyond a few weeks. They critically depend on data and gain in accuracy as more data become available to refit the curves. In that process they make strong assumptions with regards to the suitability of the data used for fitting [12]. Data can be unreliable, since the impact of interventions such as social distancing or partial lockdowns varies among geographical locations. There are many factors that determine the socio-economic costs of these measures, such as the "protection effect" and the "adjustment cost effect" [13], which can strongly influence the success of certain intervention strategies even when implemented in similar ways. Our approach neither simulates nor does it fit data. Rather, it renders a more precise characterization of local regions than other methods. It emphasizes *the explanation* of COVID-19 mortality risk, which we show has predictive value.

Other approaches use more conventional statistical models, such as correlation and linear regression to understand the influence of certain socio-economic factors, such as county-level health variables, urban density, poverty, modes of commuting, and so on, while controlling for other effects, such as race [14, 15]. Typically these results are obtained via standard step-wise modeling approaches thatare not overly scalable in the number of factors and local regions, making the discovery of significant statistical relationships rather tedious. And so, the magnitude and depth of the discoveries, while certainly insightful, have been somewhat limited. We advocate for a more comprehensive approach rooted in data science that can automatically discover precise, statistically significant relationships (called patterns) from hundreds of socio-economic features thatwould be difficult to hypothesize and test with current methodologies.

2.2 Public Information Portals

A primary source of information has been the Coronavirus Resource Center at Johns Hopkins University [1]. They constructed dashboards for the US and for the entire world that each showed the respective geographic maps overlaid with visual representations of the numbers of people tested positive alongside various test and death statics, leader boards, and temporal growth curve ensembles that compare regions at various scales in terms of the increase of test cases and deaths. Other dashboards and browser-based interactive visualization of COVID-19 related data have been made available by the companies Tableau [17], TIBCO [18], Datawrapper [19], the open source project Nextstrain [20], newspapers like the *New York Times*, and others. These dashboards and visualizations illuminate specific aspects related to the outbreak, such as race, hospital overcrowding, test statistics by state, mask compliance by county, unemployment rates and claims, effects on retail, economic inequality, pathogen evolution, and more.

Google used tracking data gathered through their Google Maps application to create what they call Community Mobility Reports [21]. These documents provide mobility trends to several types of venues, such as retail and recreation, supermarkets and pharmacies, parks, public transport, workplaces, and residential on the country, state, and county levels. The reports are helpful to assess how local communities comply with social distancing or open up economies, at least indirectly by ways of traffic load.

3 METHODS

Our study uses the prominent Kaggle repository UNCOVER COVID-19 Challenge sponsored by the Roche Data Science Coalition [22]. This dataset has about 500 attributes, or *features*, covering demographics, economics, infrastructure, and other aspects for each of the 3,007 US counties. Many of these are local socio-economic vulnerability measurements provided by the CDC [23]. The 500 features result in a 500-dimensional feature space. Our pattern mining engine looks for regions in this feature space that are occupied with similar counties that all respond in a similar way to a given target variable of interest, in this case the number of COVID-19 deaths in relation to each county's population, also called the *death rate*. Note that counties that are considered similar do not need to be geographically connected; they just need to have similar characteristics in terms of their feature values.

Each pattern of similar counties forms what is called a *subspace* [24]. It is a subpopulation of counties that fit inside a low-dimensional hypercube with well-defined value ranges of the features that describe the subspace. This property, and the fact that these subpopulations are typically rather low-dimensional, even when the overall feature space is not, makes them easy to understand and explain [25]. We exploit this property for the study presented here and note that while deep neural networks, random forests, and so on, also learn low-dimensional representations, these are not easily described in terms of their native attributes.

Concretely, given a dataset with attributes $\{A_1 A_2 \ldots . A_m\ P\}$ with P being an attribute of interest, such as COVID-19 death rate, the goal of pattern mining is to find a hypercube (or pattern) consisting of constraints of the form $A_i \in [v_l,\ v_r]$ for $i \in\ [1 \ldots m]$ (for example, *age* > 45, *race* = Asian), where the points within the pattern are "interesting." For our purposes, a pattern of counties will be considered interesting if it is associated with a COVID-19 death rate that is higher on average than the US county average. The definition of what constitutes a consistently interesting pattern is based primarily on statistical hypothesis testing. For numerical attributes, we use the Mann-Whitney test [26] to account for the often non-parametric nature of the data, while for a binary target attribute, we use the $\chi 2$ test for independence.

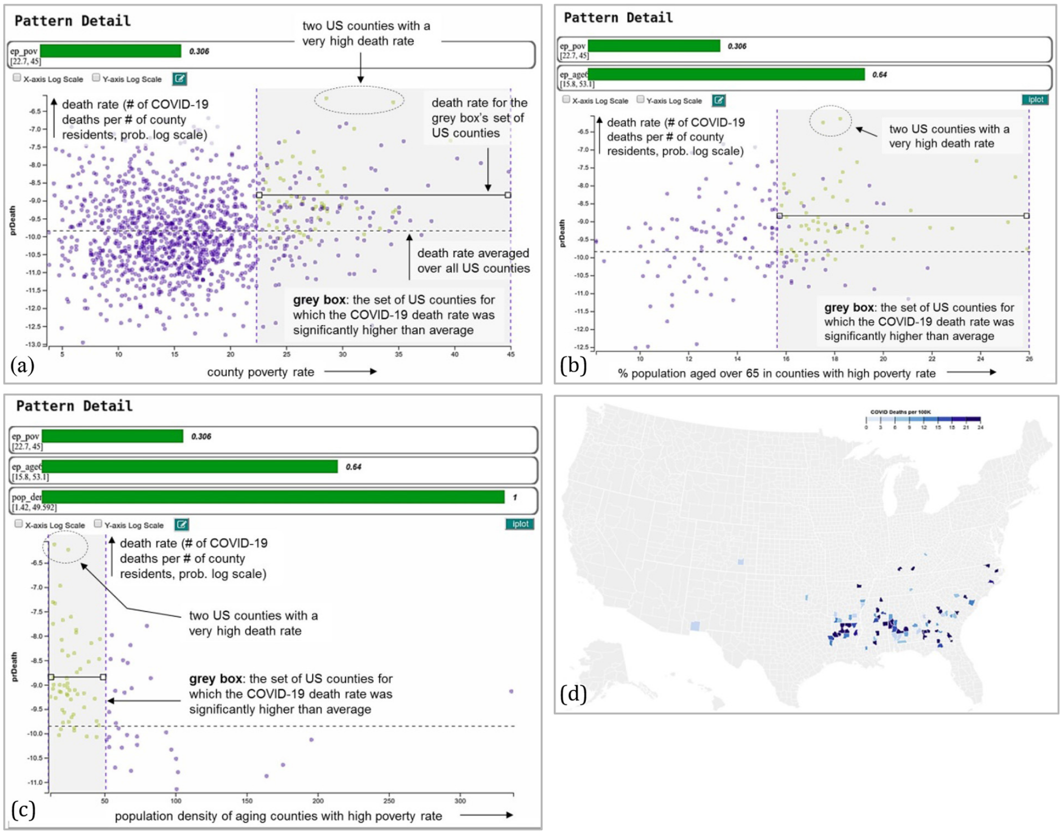

Fig. 1. Example pattern 1: (a) 1st attribute: county poverty rate; (b) 2nd attribute: age of county population; (c) 3rd attribute: county population density; (d) map with the pattern's counties, color shading denotes COVID-19 death rate.

Extracting the patterns requires extensive search. Our pattern mining algorithm is based on the FP-growth algorithm [27]. It has the nice property that it only requires scanning the full dataset twice throughout the mining process. This makes it more scalable in contrast to many other pattern mining algorithms in which the dataset must be repeatedly scanned. Given many additional optimizations, our pattern mining algorithm can analyze even large datasets within a couple of minutes.

4 FINDINGS AND RESULTS

In the following, we describe two examples of patterns we extracted and their subsequent use for predictive analysis. We also describe one additional example where we expose correlations in the extracted patterns.

4.1 Pattern 1: Poor, Aging, Rural Counties are at High COVID-19 Risk

In the visualization in Figure 1(a), each of the purple or yellow data points is a county and the placement along the horizontal axis is determined by its "poverty rate." The vertical axis denotes the COVID-19 death rate. The grey box contains the counties where on average and compared to other US counties an unusually high number of residents died from COVID-19, in relation to the county's overall population. Due to its importance, we shall call it the *Box of High Risk*. We can observe the high risk, since the points in the box tend to float in higher regions, which define higher COVID-19 death rates. The placement of the box tells us that a high poverty rate (>22.5%) is common for these counties.

We observe that the grey box has a mix of purple and yellow points; however, only the yellow points are counties where high COVID-19 death rate is consistent. The green *Bar of Confidence* on top of the plot is an important diagnostic indicator. Its length is about 30% of the full length and the number next to it reads 0.306. It means that our risk assessment is still about 70% off. Not all counties with higher poverty rate are at higher than usual risk of COVID-19 death. While we need to refine the pattern, we are on the right track, since there are only purple dots outside the grey box. This confirms that our pattern mining engine did not overlook any county that is of interest for this specific pattern description (the yellow dots).

Next, our pattern mining engine automatically refines the pattern description from above by adding “age greater than 65” as a second dimension (see Figure 1(b)). Note that we added this constraint only to the counties that were captured in the initial Box of High Risk. That is why there are fewer counties in this visualization now. It sharpens our risk assessment to counties with higher poverty rates and aging population. Fittingly, the green Bar of Confidence now reads 64%. We are getting better, but we still have more work to do to reach 100%—there are still some purple dots in the Box of High Risk.

The third automatic pattern refinement step leads to the final visualization, in Figure 1(c). It is the same set of counties as in the previous step, but now ordered from the aspect of “population density.” We observe that it is the counties with low population density thaton average have a COVID-19 death rate above US average, and our software places the Box of High Risk at the statistically correct margin. There are no more purple points in the box, which means that our risk assessment is razor-sharp. The bottom-most green Bar of Confidence confirms this; it is maxed out at 100%.

Figure 1(d) shows a map of the counties matching the pattern. Darker blue shades denote a higher COVID-19 death rate. According to the pattern’s description these are all poor and aging counties with low population density; they are on average especially hard hit by the COVID-19 virus. While it is well known by now that older residents are more vulnerable to COVID-19, the pattern tells us that this high risk seems to be amplified by two factors: (1) the residents live in sparsely populated areas thatoffer fewer urgent care facilities and (2) the residents are mostly poor, which hampers their ability to use and pay for these services.

4.1.1 Predictive Analysis. Prediction is the ultimate goal of data analytics. While this is enormously difficult and prone to many unforeseen circumstances, statistically robust pattern analysis can make predicting the future significantly more reliable. Although one can never be 100% sure, it can increase the odds to a point where it is worth taking the risk. For the particular example of epidemic spread, health officials are highly interested in identifying the US counties where the COVID-19 death rate might spike next. It would allow authorities to direct test kits, allocate hospital care, increase contact tracing, alert the community, and so on. The patterns we find are essentially predictions of the response variable—higher than average COVID-19 death rate.

In the following discussion, we turn the clock one month forward and explore whether the patterns we found are indeed able to predict a higher-than-average growth in the COVID-19 death rates. The initial analyses were conducted using data from May 10, 2020. In the following, we used the equivalent COVID-19 death statistics from June 10, 2020. On that date, the US county-wise COVID-19 death rate average was 24.1 deaths per 100K population, up from 16.8 on May 10—an increase of 7.3.

This particular pattern (poor, aging, rural counties) contains 106 counties. In May, 45 (42%) of these counties had a COVID-19 death rate greater than the then-prevailing US county average. In June this number grew to 53 (50%) with 9 counties joining this subset, and one barely leaving it. The average death rate in May was 27.6 deaths/100K, while in June it was 46.4—an increase of 18.7 (2.6 × the US average).

Figure 2(a) plots the June county death rates sorted by the May county death rates, with the respective US-wide county death rate as the zero line. Each data point along the x-axis corresponds to a county; the “county ID” is just the number determined by the May death-rate sorting. We chose a line plot over a bar chart, since it better shows

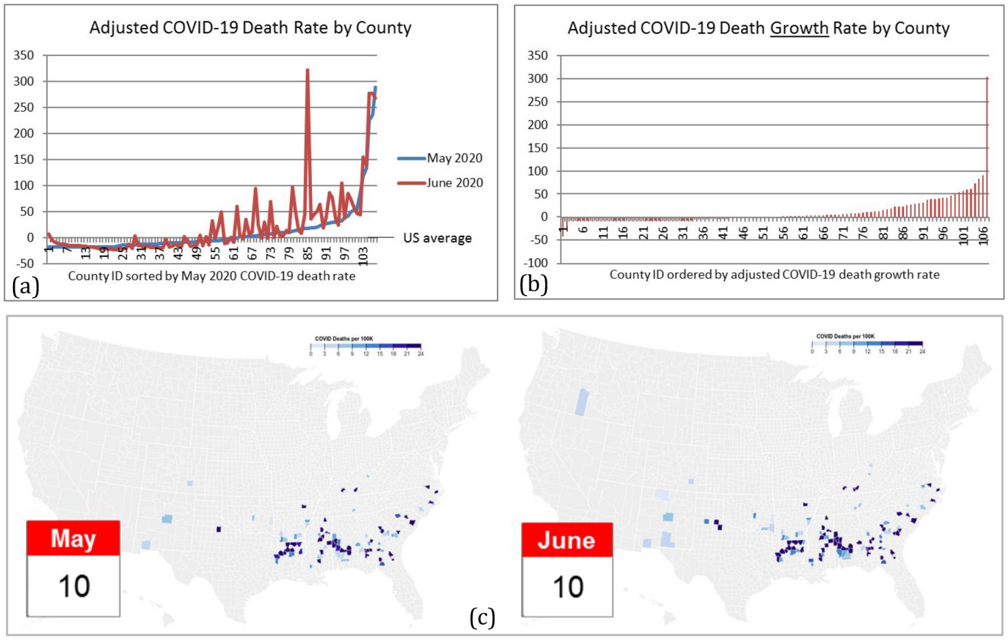

Fig. 2. Predictive power of pattern 1 (June vs. May 2020): (a) COVID-19 death rate by county adjusted for US-wide county death rate average; (b) death growth rate by county adjusted for US-wide growth; (c) county map comparison.

the peaks (each tick on the x-axis is one county). We call the plot "adjusted," since the blue curve represents the May death rates minus the US average in May, and the red curve represents the June death rates minus the US average in June. It can be clearly observed that the growth magnitudes of county death rates in June over May were vastly higher than the respective death rate declines. The pattern predicts these strong upward trends.

Figure 2(b) clarifies this even more. Here, we plot the COVID-19 growth rate for each county in the pattern, adjusted for the US-wide growth. In essence, this is the real growth generated by the pattern itself, corrected for the overall US-wide trend. We can see that while some counties down-trended slightly, a large number of them experienced strong upward growth. This stark contrast is also expressed in the two one-sided standard deviations, where the standard deviation for the counties experiencing positive adjusted growth is 43.7 deaths/100K, while the standard deviation for the counties experiencing a decline is a mere 6.0.

Figure 2(c) compares the pattern's May county map with that of June. We observe that for quite a few of these counties the COVID-19 death rate has markedly increased; the shallow blue coloring has turned to dark blue. Other counties previously not affected, but fitting the respective pattern profile, have now seen their first COVID-19 fatalities; they were invisible on the May map (colored white) but are now shaded in shallow blue on the June map.

4.2 Pattern 2: Low Asian & High Minority Population, Poor Black Kids

This pattern starts out with "percent Asian population" giving rise to a very slim Box of High Risk on the left (Figure 3(a)). The placement of the box signifies that a very low Asian population (<1.5%) is common for these counties. Given that the pattern description is not complete, our pattern mining engine automatically refines it by adding "percent minority" as a second dimension (Figure 3(b)). This sharpens our risk assessment to counties with low rates of Asians but a strong minority population (above 4%). Yet, a third refinement step is still needed, which leads to the final visualization in Figure 3(c), adding the aspect of "percentage of black children living

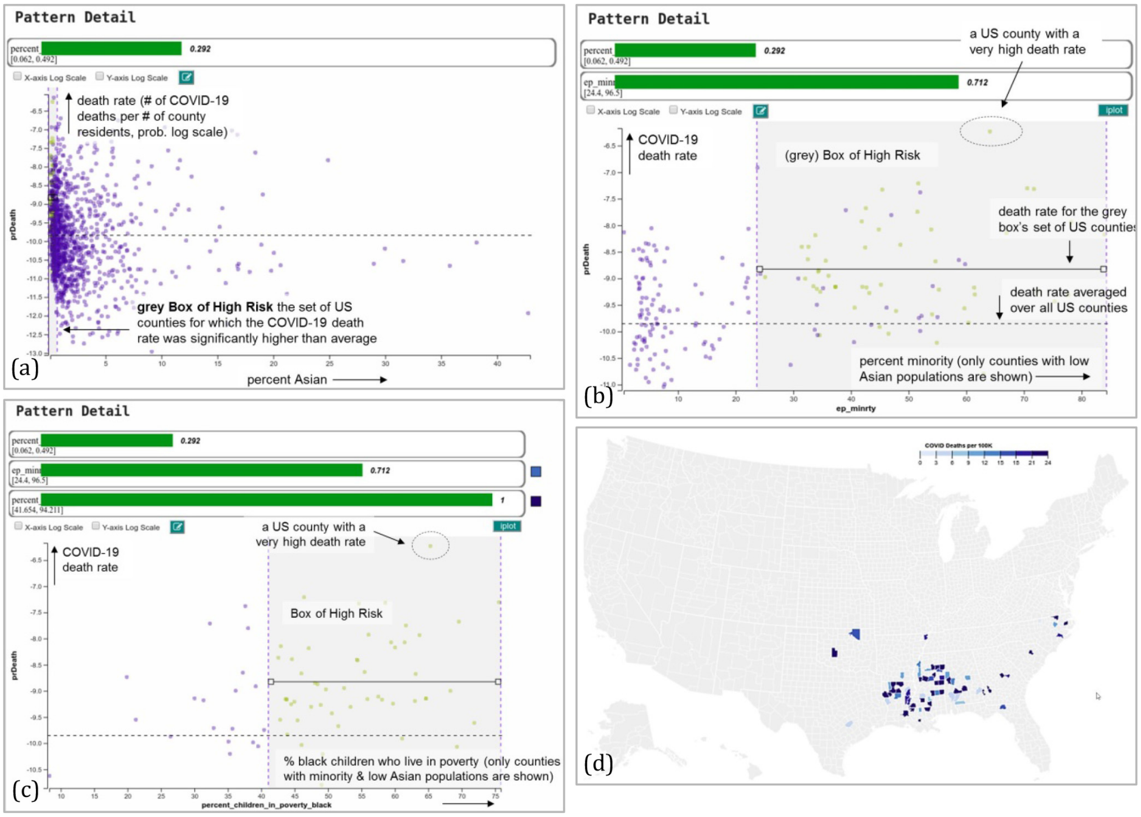

Fig. 3. Example pattern 2: (a) 1st attribute: percentage of Asian county residents; (b) 2nd attribute: percentage of minority county residents; (c) 3rd attribute: percentage of black children who live in poverty (d) map with the pattern's counties.

in poverty" to the description. We observe that counties where this percentage is above 41% have an average COVID-19 death rate above US average. A US map with these couties is shown in Figure 5(d).

This pattern reveals that counties where the minority population is high and where black children live in poverty are on average especially hard hit by COVID-19. We also learn that these counties have a very low Asian population. While the latter could indicate that awareness how to deal with a respiratory disease is low, some may say that these pattern descriptors appear somewhat over-engineered. Yet, given the complexity of the feature space this level of refinement might just be needed to separate high and low risk.

4.2.1 Predictive Analysis. This pattern is composed of 86 counties. In May, 44 (51%) of these counties had a COVID-19 death rate greater than the then-prevailing US county average; in June, this number grew to 51 (59%) with 9 counties joining this subset and 2 leaving it. The pattern's average death rate in May was 28.7 deaths/100K, while in June it was 50.0—an increase of 21.3 (nearly three times the US county-wise average).

As before, Figure 4(a) contrasts the June county death rates with those of May with the respective US-wide county death rate as the zero line. It is clear that there is a strong upward trend from May to June. Counties that already had a high death rate grew even more vigorously, while those that declined only minimally did so. Further, we can also see a significant uptick in the death rate for counties on the low end in May

The adjusted growth rate (Figure 4(b)) clarifies this even more. Only two counties (on the left side of the plot) showed a significant downward trend in the death rate. About half of the counties showed modest growths or

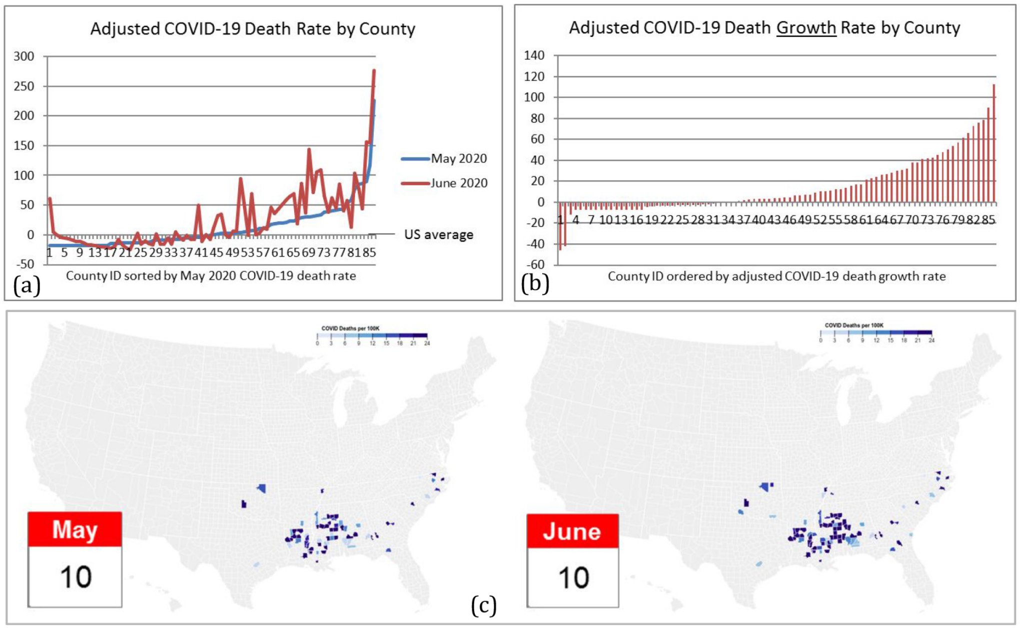

Fig. 4. Predictive power of pattern 3 (June vs. May 2020): (a) COVID-19 death rate by county adjusted for US-wide county death rate average; (b) death growth rate by county adjusted for US-wide growth; (c) county map comparison.

declines, while the other half exhibited strong or very strong growths. The standard deviation for the counties experiencing positive adjusted growth is 26.5 deaths/100K, while for those in decline it is 8.5 (32%). The high level of predictive power further confirms the choice of the three somewhat specialized pattern descriptors (see our discussion above).

4.3 Correlation Mining

Correlation can reveal a linear association between two variables, such as exercise and health. But often correlation only occurs in specific sub-populations; for example, there is a fairly strong correlation between arm length and an athlete's propensity to become a world class swimmer, while for some other sports the length of one's arm is only mildly relevant, if at all. The same applies to COVID-19 analytics. Important correlations are often hidden with conventional correlation analysis when enacted over all counties. Conversely, our pattern mining engine can reveal specific sets of counties where certain important correlations hold and so enable more targeted COVID-19 testing and health policy making. Pattern 3 examines the results of our correlation mining via one example.

4.3.1 Pattern 3: Severe Housing Cost Burden Correlates with COVID-19 Death Rates. The section headline makes this claim, but the scatterplot in Figure 5(a) suggests that there actually is no apparent correlation between severe housing cost burden and COVID-19 death rate; the correlation is a mere 10%. But in fact there are county patterns where such a correlation holds, as we will see next.

Our correlation mining revealed that there indeed is a statistically significant linear relationship between severe housing debt and COVID-19 death rate, but only for counties where the percent of home ownership is greater than 71% and where the ratio of residents living below the ethical poverty line (EPL) is less than 8%. The counties that fit this description are the purple points in the scatterplot shown in Figure 5(b). Considering only these 102 counties raises the correlation to a moderate to high level of 62%.

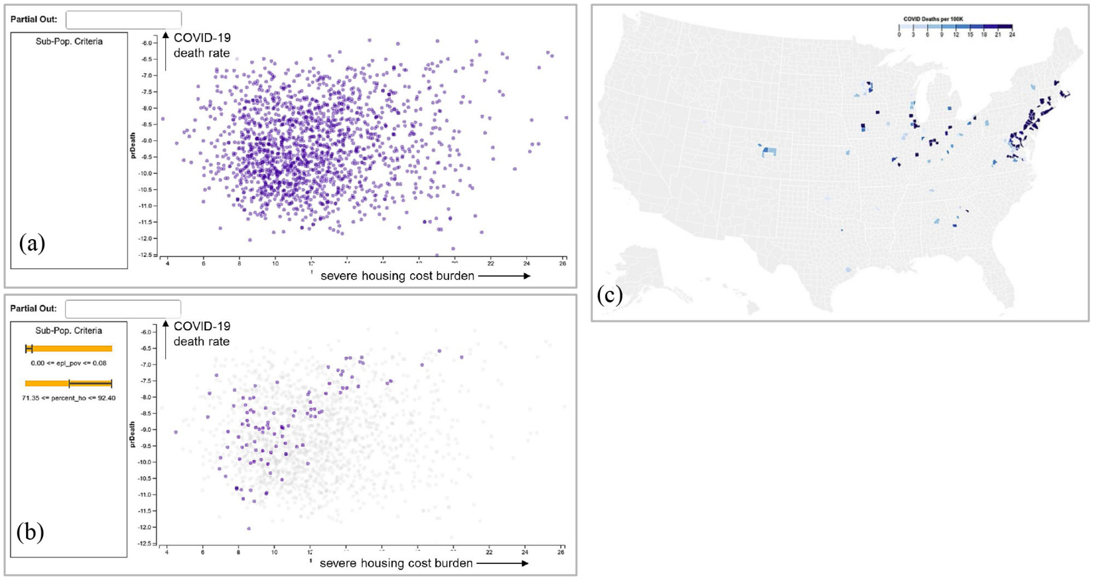

Fig. 5. Correlation mining: (a) COVID-19 death rate over the percentage of residents that suffer under severe housing cost burden for each county (ρ=0.1); (b) only for counties where the percent of home ownership was high and where the ratio of residents living below the ethical poverty line (EPL) was low. (ρ=0.62); (c) map with these counties, darker colors map to higher unemployment and COVID-19 death rate.

The pattern indicates that the residents of these counties are typically well-to-do and live in houses or apartments they own. The map in Figure 5(c) shows the counties that fit this description. They differ in the housing debt burden rate; the higher the rate, the darker the color. And due to the correlation relationship we found, darker colors also denote higher COVID-19 death rates.

We observe the higher affected (dark-blue-colored) counties are primarily in the Northeast region of the US, as well as in the large metropolitan areas in the Midwest near the Great Lakes. While these have been widely reported in the news, our analysis explains the specific characteristics that are shared among all of them, which, when coupled with high housing debt burden, can make them particularly vulnerable to COVID-19 mortality. While there are other US counties that also incurred high COVID-19 death rates, they are not characterized in this way. They have other explanations we derived as well in our studies.

As one might imagine (see also the crisis leading up to the Great Recession in the late 2000s), there are homeowners in these rich counties with high home ownership who cannot really afford their homes and as a result run high housing debt. These homeowners struggle for money and might worry a great deal about their dire situation, leading to extensive stress—a leading factor for reduced virus immune response. They may also not have the luxury of hunkering down at their house, but need to take up jobs that require them to work outside. They also might not have the funds for immediate medical care, cannot afford food delivery and possibly even a car, requiring them to take public transportation. Then, as the percentage of these types of homeowners in a county grows, so does the risk of COVID-19 infection and eventually death.

5 CONCLUSIONS AND FUTURE WORK

We have demonstrated that pattern analysis can be highly effective in defining the characteristics of counties at risk of elevated COVID-19 death rates and that these characteristics also allow reliable predictions of future death rates. The examples shown in this paper were randomly selected from a set of 297 patterns we found; we did not “cherry pick” the best results (more examples are on our webpage [28]). Overall, we compared the death rates in May with those in June and found that for 98% of our patterns the death rate growth was 2–3 times higher than the US average, while the remaining 2% grew at the average pace, and none slowed in growth below the US average. These trends continued in July as well.

It is important to note that our patterns also identified counties that fit a certain profile but where the spread of the disease had not hit yet. Others grew even more. Recognizing these trends early can be of tremendous help in planning and advocating for the appropriate set of resources to ease the impact of the disease.

There are other machine learning approaches such as neural network, random forest, decision trees, or the like that could be used to make predictions. But these predictions would just be probabilities of risk or death rate figures. Health officials would not know *why* the risk exists or *why* the death rates will be so high. This is a downside of what has become known as “black box AI.” The advantage of our explainable AI approach is that it can explain why a county is at risk in the language of the factors that expose this risk. We believe that these explanations and the stories they tell can help human decision makers in gaining trust in these predictions, especially in a culture where predictions are often discounted at a local level. Furthermore, health officials will also better understand why certain virus spread interventions might work better than others, and argue for these more successfully in their local communities.

Current work focuses on providing recommendations for interventions. We can take a certain mitigation measure as the target variable and use our pattern mining algorithm to identify groups of counties where the outcome of this measure (assessed in decline of death rate) was higher (or lower) than usual.

REFERENCES

[1] J. Brainard. 2020. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?” *Sci. Mag. Online.* May 13, 2020. Retrieved from https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat.

[2] 2020. CORD-19: COVID-19 Open Research Dataset. Retrieved from https://www.semanticscholar.org/cord19.

[3] 2020. SSRN Coronavirus and Infectious Disease Research. Retrieved from https://www.ssrn.com/index.cfm/en/coronavirus/.

[4] A. Bertozzi, E. Franco, G. Mohler, M. Short, and D. Sledge. 2004. The challenges of modeling and forecasting the spread of COVID-19. *ArXiv Preprint ArXiv*:2004.04741.

[5] C. Hou, J. Chen, Y. Zhou, L. Hua, J. Yuan, S. He, and Y. Guo, et al. 2019. The effectiveness of quarantine of Wuhan City against the corona virus disease 2019 (COVID-19): A well-mixed SEIR model analysis. *J. Med. Virol.* 2020.

[6] Z. Yang, Z. Zeng, K. Wang, S. Wong, W. Liang, M. Zanin, and P. Liu, et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* 12, 3 (2020), 165.

[7] 2020. University of Washington Institute for Health Metric and Education (IHME) COVID-19 Resources. Retrieved from http://www.healthdata.org/covid.

[8] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, and T. Yang, et al. 2020. Predicting covid-19 in China using hybrid AI model. *IEEE Trans. Cybern.* 2020.

[9] S. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. Atkinson. 2020. Covid-19 outbreak prediction with machine learning. *SSRN Preprint*, 2020 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3580188.

[10] P. Bruce. 2020. Covid-19: epidemiological models vs. statistical models. *Elder Res.: Data Sci., Mach. Learn., AI*, April 13, 2020. Retrieved from https://www.elderresearch.com/blog/covid-19-epidemiological-models-vs.-statistical-models.

[11] I. Holmdahl and C. Buckee. 2020. Wrong but useful—what Covid-19 epidemiologic models can and cannot tell us. *New England J. Med.* 2020.

[12] N. Jewell, J. Lewnard, B. Jewell. 2020. Caution warranted: Using the institute for health metrics and evaluation model for predicting the course of the COVID-19 pandemic. *Ann. Intern. Med.* 173, 3 (2020), 226–227.

[13] C. Brow and M. Ravallion. 2020. Inequality and the coronavirus: Socioeconomic covariates of behavioral responses and viral outcomes across US counties. *National Bureau of Economic Research Working Paper*, No. 27549, 2020. Retrieved on August 1, 2020 from https://www.nber.org/papers/w27549.

[14] C. Knittel and B. Ozaltun. 2020. What does and does not correlate with COVID-19 death rates. *National Bureau of Economic Research Working Paper*, No. 27391, June 2020. Retrieved from https://www.medrxiv.org/content/10.1101/2020.06.09.20126805v1.

[15] F. Carozzi. 2020. *Urban Density and Covid-19.* Research Paper. Institute for the Study of Labor (IZA). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3643204.

[16] Coronavirus Resource Center at Johns Hopkins University. Retrieved from https://coronavirus.jhu.edu/.

[17] Coronavirus Global Data Tracker. Retrieved from https://www.tableau.com/covid-19-coronavirus-data-resources.

[18] Coronavirus Visual Analysis Hub. Retrieved from https://www.tibco.com/covid19.
[19] Datawrapper interactive visualization. Retrieved from https://www.datawrapper.de/.
[20] Nextstrain real-time tracking of pathogen evolution. Retrieved from https://nextstrain.org/.
[21] Google COVID-19 Mobility Reports. Retrieved from https://www.google.com/covid19/mobility/.
[22] UNCOVER COVID-19 Challenge data. Retrieved from https://www.kaggle.com/roche-data-science-coalition/uncover.
[23] The CDC Social Vulnerability Index. Retrieved from https://svi.cdc.gov/.
[24] H. Kriegel, P. Kröger, and A. Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* 3, 1 (2009) 1.
[25] B. Wang and K. Mueller. 2014. Does 3D really make sense for visual cluster analysis? Yes! In *Proceedings of the IEEE VIS Workshop on 3DVis: Does 3D Really Make Sense for Data Visualization?*.
[26] P. McKnight and J. Najab. 2010. Mann-Whitney U test. *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc.
[27] J. Han, J. Pei, and Y. Yin. 2000. Mining frequent patterns without candidate generation. *ACM SIGMOD* 29, 2 (2000), 1–12.
[28] COVID-18 Data *Analytics*. Retrieved from https://akaikaeru.com/covid-19-1.

Received June 2020; revised August 2020; accepted October 2020