Impact of COVID-19 on the US Housing Market

Jianan Song, Chao Wang, Yifan Tang, Jiebei Luo, Po-Han Lai, Mingzhi Dai

1 INTRODUCTION - MOTIVATION

In 2020, a new respiratory disease, commonly known as the coronavirus disease (COVID-19), gripped the world([1] and [2]). Despite worldwide government shutdown orders, and practice social distancing, the disease inevitably triggered recession ([16], [7], [13]).

The pandemic directly impacts the economy through sharp decreases in consumer spending ([9], [15]) increase in unemployment ([3], [5], [8]), and volatility in equity returns ([17]). [8] shows that this pandemic disproportionately affected older workers of ages 65 and older, particularly for women, whose unemployment rate is significantly higher than men across all age or ethnicity groups. Meanwhile, a rapidly growing literature on the economic effects of investigate whether recent policy interventions have been successful ([6], [19], [16]). Housing related issues attracted immediate attention with shelter-in-place orders, self-isolation, and quarantine as the primary global responses to the COVID-19 pandemic,showing that countless people are struggling paying for house mortgages or rents due to the heterogeneous impact on industries and areas brought by the pandemic. With the pandemic, housing is more than a place of residence – it becomes an office, a classroom, and a shelter ([17] and ([12])). It would be particularly interesting to study how housing market has been affected by coronavirus and make predictions on how it may evolve in the near future.

We would like to know if uncertainty situations of COVID-19 affect the house price, house value and house inventory, while most literature about the socio-economic consequences of COVID-19 focus on the labor market, health outcomes, gender and racial inequality, and environmental outcomes, the consequences of COVID-19 on housing market is almost an unexplored topic.

Among literature about effects of COVID-19 on the housing market, most of those articles either don't have much data of the correlations between housing market and COVID-19 or the data is only from one or two month time period ([2], [10] and [4]). Not to mention, there was very limited data visualizations to directly show the relations of the two from different perspective, especially when the housing market measures are limited to housing with federally backed mortgages ([18], [14] and [11]).

2 METHODS & APPROACH

2.1 Intuition

Most literature apply difference-in-difference methods to estimate the effect of COVID-19 on housing market based on the various shut-down timeline in each state. We predict how the housing market perform during COVID-19 based on the housing data in the past three years. The innovation aspect of our approach is to train our housing market model based on housing data prior to COVID-19 and use the trained model to predict the housing performance during COVID-19.

In our study, we implement various machine learning techniques including K-means model, ARIMA, LSTM, RNN for predicting and clustering. In clustering, gap-statistics are utilized to determine the K with higher confidence level. In data visualization, we use parameters as the filter that enables a dynamic column switching between parameters. Additionally, we create an interactive dashboard to show the animation on how the COVID-19 situations have developed and the housing clusters have evolved, and then how the relationship between the COVID-19 risk and house total values have changed over time interactively.

2.2 Prediction Models in Housing

We predict how the housing market perform during COVID-19 based on the housing data from 2017 to 2019. Our approach is to innovatively train our housing market models with the housing data prior to COVID-19 and the use the trained models to predict the housing performance during COVID-19. The predicted data is compared with the actual data during COVID-19 to quantitatively visualize the impact of COVID-19.

The objective is to predict the total sold value in the housing market since the COVID-19 started, thus we choose ARIMA for time series analysis. ARIMA is a generalization of an autoregressive moving average

(ARMA) model and widely applied in cases where data show evidence of non-stationarity.

To start with, we first run time series model, autoregressive integrated moving average model (ARIMA), as our baseline model and then improve them using Seasonal ARIMA (SARIMA) and Long Short-term memory (LSTM) models.

Steps applied in the ARIMA model are shown in Figure 1. Tune AR (Auto Regressive), I(difference), MA (Moving Average) parameters which are denoted as p, d, q. The initial values are obtained using exploratory data analysis on the sold housing value. For the parameters AR and MA (p and q), we use partial autocorrelation (PACF) plot and autocorrelation plot to decide initial values. For the difference (parameter d), we use Augmented Dickey Fuller test which is a common statistical test used to test whether a given time series is stationary or not. 2. After having initial values for AR, I and MR (parameters p, d, and q), we apply the grid-search to find the optimal values for the AR and MR combinations. We execute the search by adding one unit for each step and there are 9 possible combinations in our project. The best ARIMA performed model is determined by the minimum Root-mean-square deviation (RMSE) on the training dataset.

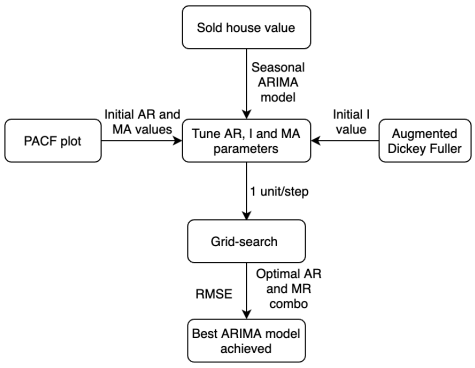


Figure 1: ARIMA work flow

After exploratory data analysis, we find that the house total value has seasonal change, thus we apply Seasonal ARIMA to account for the seasonal effects in the housing market. We set d = 1, the degree of differencing, and start p and q from 0 with function autoarima from package pmdarima. In terms of season factor m, we use m=52, because our data set is by weekly. In this function, we use the lowest AIC to select the best p,q and seasonal P and Q. Moreover, we plan to apply the Long Short-term memory (LSTM) model of the recurrent neural networks (RNN). RNN is a powerful type of neural network designed to handle the sequence dependence since it persists all previous information as a short-term memory and make decisions based on all stored information in the LSTM cells. In this project, we use the Vanilla Long Short-term memory (LSTM) and the detailed steps are shown in Figure 2.

1. Transfer our time series data into a supervised learning problem with input as X and output as Y, which is how Pytorch implements the LSTM model. Specifically, we use the house value in t to t+51 period as our input X and the house value in t+52 period as our output Y. Besides, we use the MinMaxScaler function in the sklearn package to transform the observations to have a specific scale since the neural networks expect data to be within the scale of the activation function.

2. After the data processing, we first fit our model using the Vanilla LSTM, which has a single hidden layer of LSTM units to forecast more complex data trends. For each LSTM layer, we construct inputs as samples and feature. Specifically, for our case, samples are our rows of house data. There is only one feature as we mainly use total sold housing value.

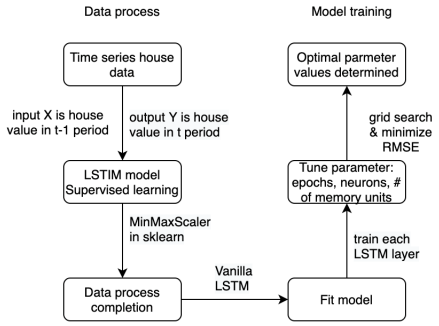


Figure 2: LSTM work flow

2.3 COVID-19 Clustering

To evaluate the housing market performance during COVID-19, we cluster states into groups and investigate the effect on the group level. Clustering algorithm is an unsupervised Learning type.

For clustering that offers excellent data visualizations, K-means can be used to find states that may experience similar levels of COVID-19 impacts. Our clustering features include two COVID-19 metrics: new cases and

cumulative cases. The two features are normalized into the same scale within 0 to 1 under the assumption that each feature is equally important.

To perform K-means clustering at the state level, we randomly pick K centers and assign the states to each group according to the similarity defined as the distance. We then recalculate the centers (the average of all data in each group) given the grouping results. We keep updating the grouping and centers until converge. We define converge when all states are assigned to one of the K groups and the overall within-group variation is minimized. This overall variation is the sum of the distances between each data point and its assigned center.

Next, gap-statistics is used to determine the K, which compares the difference between the overall variation from our data (all states) with the variation data (i.e., same number of data with random values). Then we test out all K numbers and check the gap-statistics, the highest difference is associated with the right number of K.

The same clustering steps is applied to the monthly COVID19 data. Lastly, we use the most common K as the final K values to generate the clustering results for each month. In order to better interpret the clustering result, we calculate the center of each cluster and compare with the percentile of the case distribution to determine if the center of cluster is low or high. The description of each group can be interpreted as: cluster 1, high risk states where both new and cumulative cases are high; cluster 2, medium to low risk where new and cumulative cases are in medium or low range.

2.4 Visualizations

Tableau is employed in our project as the primary visualization presentation platform. By creating a story, our project incorporates three units that clearly narrate how the home value in the U.S. has been affected by the COVID-19 pandemic. Essentially, the descriptive statistics of the U.S. housing market (2017-2020) the COVID-19 data, and the analysis results of our model are realized by utilizing multiple visualization tools in Tableau with interactive features to demonstrate the impact of COVID-19 on the U.S housing marking in the past 10 months.

2.4.1 Visualization of U.S housing market. A line plot is chosen to demonstrate changes in home sales. By plotting two outcome variables, i.e., total homes sold and median sales price, against the time variable (day of the year), the line plot displays the year-by-year growth patterns of the U.S. housing market in the past 4 years. Secondly, filters of year, homes sold in the past 4 and 12 weeks, and median sales price in the past 4 and 12 weeks are also added to improve the flexibility of using the dashboard.

2.4.2 Visualization of COVID-19 cases and deaths. The COVID19 case and death information is also respectively visualized on a U.S. map (by state). By applying the color to the number of cases (and deaths), the outbreak of this pandemic is displayed side by side in two different colors. More importantly, the case and death maps are utilized as a filer for the case and death bar charts, respectively. Thus, the time series visualization of COVID-19 data can be further realized at state-level. By connecting the maps with bar charts, the dashboard users are able to have the flexibility of combining multiple selection dimensions, including geographic level, time period as well as COVID-19 data on new cases and deaths to better understand the development of this pandemic.

2.4.3 COVID-19 Case Clustering and U.S. Housing Market Prediction. We combine the clustering results with state-level heatmaps of COVID-19 cases to interactively show the impact of COVID-19 on the housing market with animation effects. 50 states are divided into high risk and med/low-risk categories, with different colors reflecting the severity based on the number of new cases. A small combo chart with line and bar charts jointly demonstrating the number of new cases and cumulative cases is added to the dashboard to provide yet another dimension of COVID-19 outbreak.

In the second part, the impact of COVID-19 on the home values is visualized in a heatmap as well. The innovative approach in our visualization is that the color scale is determined by the percentage of change between the actual values and our predicted values (absent of the pandemic), computed based on the total number of homes sold and the median sales price as well as the results obtained from the ARIMA/LSTM models. The percentage change difference is then realized via the color function and can vary by month. Similarly, a scatter plot is added as a secondary visualization in this part, which plots the total home values and new

COVID-19 cases in two risk clusters by month. By applying the animation to the same month variable, a clear trend difference between high risk and medium-to-low risk clusters can also be demonstrated dynamically by month.

3 EXPERIMENTS/EVALUATION

3.1 Description of Testbed

In order to find the model with the minimal error in predicting the total sold volume, we use the value as baseline to see how the housing market deviate from the baseline under the influence of COVID-19.
In clustering, we seek to find the clustering result with the labels that can best describe the different underline groups on COVID-19 monthly data. We expect the clustering result can show the relationship between COVID-19 and housing market change differently in each group.

3.2 Dataset Preparation

Dataset 1 is the housing market data by Redfin[1]: It provides housing market daily data for metropolitan areas, cities, neighborhoods and zip codes in the US. It includes more than 30 housing metrics(columns) with a combined data size 800MB that covers the data from 2017 to 2020. Dataset 2 is the COVID-19 data released by NY Times on Github[2]. This dataset shows daily cumulative counts of coronavirus cases in the US, at the state and county level. Dataset 3 is populated (inner join) from housing data and COVID-19 data discussed above with time frame March to October 2020.

Python Pandas package is used for data processing and cleaning. Keras, Statsmodels are utilized for housing market predition. Skelearn is applied for clustering construction. The team use Jupyter Notebook for coding interactively.

3.3 Prediction Evaluation

The data is split into three sub-datasets, including training, validation and test data. Training data is designed for training the ARIMA, SARIMA, and LSTM models. Validation data is created for model comparisons, test data is reserved for comparing with predictions and visualizations. 03/2020 is the split point for test data. We use 80% data before 03/2020 for training (01/2017 - 07/2019), and 20% for validation (08/2019 - 02/2020). Then, we combine train and validation data to train models and predict the values based on the test data after comparing the results obtained.

We apply the ARIMA model by preparing and tuning parameter on a training data and evaluating predictions with the validation data. Next, the prediction is conducted on the validation data by taking a time series dataset as input and a tuple with the p, d, and q parameters required by the model, and calculate error score for predictions towards expected values. The combination of p, d and q with lowest RMSE is used to finalize the ARIMA model. Each state has different p,d and q, which are used to predict the total house value after COVID-19.

In LSTM modelling, learning rate = 0.01, windows = 52 and hidden layer size =200, which represent 200 neurons in one layer as input values. In terms of loss function, we use MSELoss function from pytorch packages to train our model and find out the better model. similarly, the LSTM model with lowest RMSE in validation data is our final LSTM model.

Based on the results from two models, we find that seasonal ARIMA outperform ARIMA so that we use SARIMA to compare with the LSTM models.

3.4 Clustering Evaluation

Gap statistics is used to determine the K, which compares the total within-cluster variation for different values of k with their expected values under the null reference distribution of the data (i.e., a distribution with no obvious clustering). In our context, it compares the difference between the overall within-cluster variation among all states with the variation from the data with random values. Next, we test out all K values and check the difference (Gap-statistics) to identify the right number of K according to the highest difference.

Next, we repeat the same steps for each month. Given that the optimized K in each month could be different, we adopt the most common K as the final K value to generate the clustering results for each month. Once state-level groups are obtained with the grouping arrangement, a label is create based on the center value in each cluster to show if the value is low or high. The labels of groups could be interpreted as: cluster 1, high risk states where both new and cumulative cases are

[1]https://www.redfin.com/news/data-center/
[2]https://github.com/nytimes/COVID-19-data

high; cluster 2, new and cumulative cases are in medium or low range. The final clustering is applied in our visualization, namely, in Section 2.4.3.

3.5 Survey

During the process of designing visualization dashboard, a survey has been conducted to solicit feedback and suggestions. Particularly, based on audience's perspective, we refer to the housing market trend graph and build the initial version of line plots and share among our team members and their friends.

4 OBSERVATIONS

4.1 Housing Market & COVID-19

Tableau dashboard click here: Interactive Dashboard As visualized in Figure 3, the line plot of total homes sold displays nearly identical trends from 2017 to 2019. Yet, the total houses sold in the year of 2020 presents a drastically different trend. Specifically, our visualization captures a clear drop-off from March to May. But after hitting the lowest point, the housing market starts to bounce back and actually surpassed the average growth trend in the past three years during the same season, but a simple visualization of COVID-19 cases and deaths in the U.S. (described in Section 2.4.2) does not demonstrate a clear pattern that can provide direct evidence toward explaining such abnormal fluctuations in the housing market in the past 10 months.

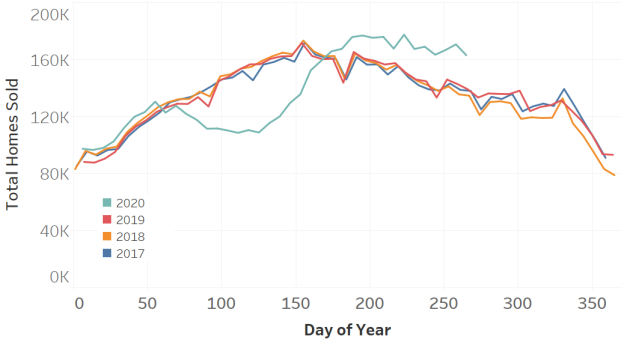


Figure 3: Total number of homes sold 2017-2010

4.2 COVID-19 Clustering

The clustering that divides 50 states into the high-risk and mid/low-risk clusters.Specifically, the centers of the high risk states and mid/low-risk states are (New Cases$_{Mean}$ = 63,308, New Cases$_{stdev}$ = 34,842, Cumulative New Cases$_{Mean}$ = 209,890, Cumulative New Cases$_{stdev}$ = 128,930) and (New Cases$_{Mean}$ =10,327, New Cases$_{stdev}$ = 5,396, Cumulative New Cases$_{Mean}$ = 38,686, Cumulative New Cases$_{stdev}$ = 29,210). By moving the cases bar through the selection of different risk clusters, we observe a contrasting COVID-19 development trend between these two clusters especially during the first three months of the pandemic. In particular, the high risk cluster encounters a quick rise and dropoff during this period (as shown in Figure 4) but the mid/low-risk cluster does not exhibit such pattern early on in the pandemic (as shown in Figure 5).

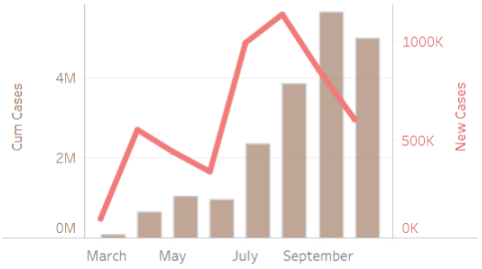


Figure 4: COVID-19 cases at high-risk cluster

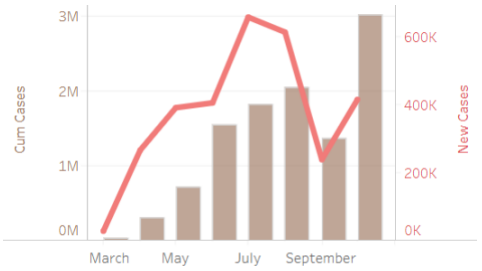


Figure 5: COVID-19 cases at mid/low-risk cluster

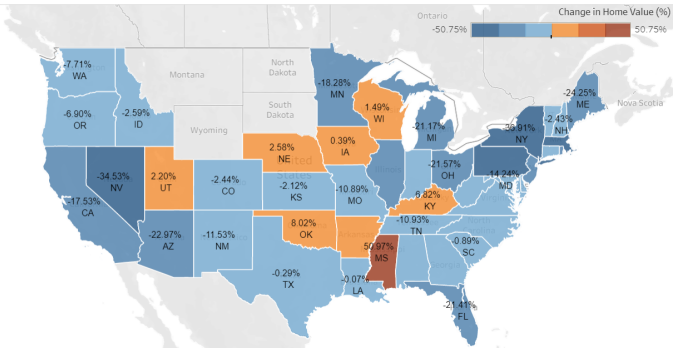
4.3 Housing Market Prediction

We fit the housing market of 50 states using different models respectively, i.e., LSTM and SARIMA. Then we select the model with lower root-mean-square error (RMSE) in the validation data, particularly, 33 states (AL, AZ, CA, CO, DC, DE, FL, ID, IN, KS, KY, LA, MA, MD, MN, NC, NH, NJ, NV, NY, OH, OR, PA, RI, SC, TN, TX, VA, VT, WA, WI, WV) use SARIMA to predict the future housing value while 14 states (AK, AR, CT, GA, HI, IA, IL, ME, MI, MO, MS, NE, OK, UT) use the LSTM

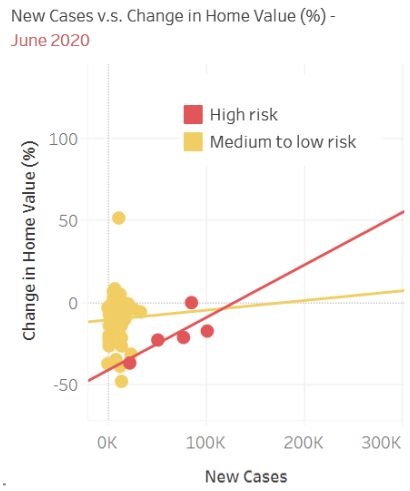
model. Overall, SARIMA model demonstrates a better performance in more states according to our analysis results. Given that the period between January 2020 and February 2020 are hard to predict in most states due to the sharp drop in the total house values, we use 52 previous data points to predict the future value.

4.4 Housing Market with COVID-19

By utilizing clustering, the US states are divided into high-risk and mid/low-risk groups by the monthly COVID-19 case data. A bigger reduction in housing activity within states in the high-risk group is detected at the beginning of the sample period (i.e., March 2020 - May 2020), but quickly catches up without being affected by the quickly growing COVID-19 cases. There is a special case in NY state, which shows a much slower recovery and we believe it is related to the local policy which has much longer shut-down compared to other states. For mid-low risk states, the effects brought by COVID-19 appear to be more mild which see less housing market activities reduction at the beginning of COVID-19 period but demonstrate a more smooth recovery. As Figure 6 shows in June 2020, COVID-19 has a significantly positive impact on the housing value in high-risk areas while the impact magnitude for the mid/low-risk areas is much flatter.



(a) Change across US states



(b) Change with clusters

Figure 6: Sold home value change June 2020

5 CONCLUSIONS & DISCUSSION

Based on our analysis, we uncover the impact differences of COVID-19 among states in U.S., in particular, we found that the housing market is more active in the high-risk areas both in sales volume and total home values. More importantly, we predict that the total housing value and housing market activities in the next three month continues to grow at a higher speed compared to the states in the mid/low risk cluster.

Meanwhile, limitations also exist in our study. Firstly, factors related to population, interest rate and government policy that can jointly impact the housing market are not included in our model. For prediction, due to the difficulty of locating a parameter for prediction in each state, performance of the LSTM model is less satisfied since we run out the degree of freedom to tune the parameters. Also, states with housing market activities have larger variations and the prediction performance is not good because of higher RMSE values in the validation data. In future studies, we consider increasing layers for the LSTM model from one to more and use more variable as input features to improve the accuracy and applicability our this model.

6 DISTRIBUTION OF TEAM MEMBER'S EFFORT

(1) Data process cleaning: CW, JL (2) EDA on the COVID-19 and interactive map dashboard: JL, JS, MD (3) ARIMA models and k-means clustering: PL, YT, CW (4) Clusters analysis based on the COVID-19 data: YT, CW - 2 wks. (5) Study survey: PL, JS, MD. (6) Data visualization: YT, JL, PL. (7) Final Report and poster: all. Every member has contributed to this project equally.

REFERENCES

[1] Dave Altig, Scott Baker, Jose Maria Barrero, Nick Bloom, Phil Bunn, Scarlet Chen, Steven J Davis, Julia Leather, Brent Meyer, Emil Mihaylov, et al. 2020. Economic uncertainty before and during the COVID-19 pandemic. *Journal of Public Economics* (2020), 104274.

[2] Mariana C Arcaya, Yael Nidam, Andrew Binet, Reann Gibson, and Vedette Gavin. 2020. Rising home values and Covid-19 case rates in Massachusetts. *Social Science & Medicine* (2020), 113290.

[3] Louis-Philippe Beland, Abel Brodeur, and Taylor Wright. 2020. COVID-19, stay-at-home orders and employment: Evidence from CPS data. (2020).

[4] Nicholas Biddle, Ben Edwards, Matthew Gray, and Kate Sollis. 2020. COVID-19 and mortgage and rental payments: May 2020. *ANU Centre for Social Research and Methods, Canberra, June, viewed* 8 (2020).

[5] George J Borjas and Hugh Cassidy. 2020. *The adverse effect of the covid-19 labor market shock on immigrant employment*. Technical Report. National Bureau of Economic Research.

[6] Mike Brewer and Laura Gardiner. 2020. The initial impact of COVID-19 and policy responses on household incomes. *Oxford Review of Economic Policy* 36, Supplement_1 (2020), S187–S199.

[7] Abel Brodeur, David M Gray, Anik Islam, and Suraiya Bhuiyan. 2020. A Literature Review of the Economics of COVID-19. (2020).

[8] Truc Thi Mai Bui, Patrick Button, and Elyce G Picciotti. 2020. *Early Evidence on the Impact of COVID-19 and the Recession on Older Workers*. Technical Report. National Bureau of Economic Research.

[9] Raj Chetty, John N Friedman, Nathaniel Hendren, Michael Stepner, et al. 2020. *How did covid-19 and stabilization policies affect spending and employment? a new real-time economic tracker based on private sector data*. Technical Report. National Bureau of Economic Research.

[10] Vincenzo Del Giudice, Pierfrancesco De Paola, and Francesco Paolo Del Giudice. 2020. Covid-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy). *Social Sciences* 9, 7 (2020), 114.

[11] Augusto Ricardo Delgado Narro and Yuya Katafuchi. 2020. COVID-19, state of emergency, and housing market. (2020).

[12] Walter D'Lima, Luis A Lopez, and Archana Pradhan. 2020. Covid-19 and housing market effects: Evidence from us shutdown orders. *Available at SSRN 3647252* (2020).

[13] Nuno Fernandes. 2020. Economic effects of coronavirus outbreak (COVID-19) on the world economy. *Available at SSRN 3557504* (2020).

[14] Laurie Goodman and Dan Magder. 2020. Avoiding a COVID-19 Disaster for Renters and the Housing Market. *Washington, DC: Urban Institute* (2020).

[15] Akos Horvath, Benjamin Kay, and Carlo Wix. 2020. The COVID-19 Shock and Consumer Credit: Evidence from Credit Card Data. *Available at SSRN 3613408* (2020).

[16] Peterson K Ozili. 2020. Covid-19 pandemic and economic crisis: The Nigerian experience and structural causes. *Available at SSRN 3567419* (2020).

[17] Peterson K Ozili and Thankom Arun. 2020. Spillover of COVID-19: impact on the Global Economy. *Available at SSRN 3562570* (2020).

[18] Yichen Su and Sitian Liu. 2020. The Impact of the COVID-19 Pandemic on the Demand for Density: Evidence from the US Housing Market. *Available at SSRN 3661052* (2020).

[19] Rob Wallace, Alex Liebman, Luis Fernando Chaves, and Rodrick Wallace. 2020. COVID-19 and Circuits of Capital. *Monthly Review* 72, 1 (2020), 1–13.