The Effect of COVID-19 on Gig Workers' Tips: Evidence from Gopuff

Srikar Katta
srikar.katta@duke.edu
Regina Ruane
ruanej@wharton.upenn.edu

Abstract

Gig workers have become a vital component of America's labor force. Understanding how disasters – such as COVID-19 – affect compensation for gig workers is essential, especially because gig workers typically do not qualify for employer-sponsored benefits like sick leave or insurance. We investigate whether individuals will increase prosocial behavior to gig workers during times of crisis in the context of the COVID-19 pandemic, delivery drivers, and tips using data from the e-commerce platform Gopuff. Employing a model-based counterfactual imputation method, we estimate that COVID-19 caused an 8% increase in tips given to drivers.

Keywords: Gig economy, COVID-19, Tipping, Prosocial behavior, Disasters

1 Introduction

The gig economy is an economic structure in which platforms, typically online, match independent contractors – known as *gig workers* – to fulfill consumers' requests. Gig workers have become a vital component of the US labor force. In 2015, there were 68 million gig workers, and over 40% of those people used gigs as their primary source of income Manyika et al., 2016. Research into the gig economy has forecast increased demand for gig workers Vallas, 2019 and even the end of traditional employment Sundararajan, 2017, highlighting the importance of the gig economy and the gig workforce.

While the gig workforce has been well researched, a prevailing issue is how economic, natural, and/or social crises – such as wars, hurricanes, and pandemics – impact compensation practices in the workforce. Between 2010 and 2019, there were over 120 natural disasters that cost the US government upwards of one billion dollars each "U.S. Billion-Dollar Weather and Climate Disasters", 2022. Additionally, Russia's invasion of Ukraine shook economies globally, evidenced by the price of oil increasing by 1.5 times Liadze et al., 2022. And of course, COVID-19 affected anything and everything from the environment Girdhar et al., 2021 to the economy Sharif et al., 2020. While traditional employees may receive sick leave or insurance that may be incredibly important during times of crises, gig workers typically are not afforded the same benefits Bajwa et al., 2018. Therefore, understanding the following is essential for characterizing the strength of the gig workforce in the future: will individual customers change their prosocial behaviors to aid gig workers who are victims during times of crises? Otherwise, will government interventions be necessary to help aid gig workers?

We explore this question in the context of COVID-19's effect on customer's prosocial behavior towards gig workers, as measured by customers' tipping practices, using data from the e-commerce platform Gopuff. Gopuff is a digital delivery service that utilizes gig workers to instantly deliver household items from warehouses to customers' front doors. We accessed data from 50,000 customers and over 600,000 transactions between May 2019 and May 2020 to characterize the causal effect of COVID-19 on tips given to delivery drivers.

Most observational techniques in causal inference depend on the existence of a set of units untreated by the intervention (i.e., a control group) Abadie et al., 2010; Card and Krueger, 1993; alternative techniques take advantage of relationships between related time series to estimate a causal effect Brodersen et al., 2015. However, COVID-19 impacted everything everywhere, therefore invalidating such techniques. We adopt a model-assisted approach to impute a counterfactual and measure the effect of COVID-19 on tipping over time.

After curating, filtering, and cleaning the data, we use tipping data from seven metropolitan areas to conclude that tipping behavior increased by approximately 8% after COVID-19. These findings suggest that gig workers received some benefits from the public and individuals; however, an 8% increase in tips may not account for the hazards associated with working in the field during COVID-19.

2 Related literature

2.1 Prosocial Behavior and Tipping

Prosocial behavior related to the pandemic has become increasingly important, as evidenced by the depth of research in the area. For instance, Rieger et al., 2020 found that triggering altruism – a key prosocial behavior akin to selflessness – leads to a greater willingness to be vaccinated. Additionally, Alfaro et al., 2020 found that stringency measures matter less in more altruistic communities, suggesting that increased altruism may indicate more relaxed government regulations. Taken together, these studies suggest that changes in altruism can redefine government responses to the pandemic.

A key concern for measuring prosocial behavior is the lack of a viable proxy. While some studies utilize survey methods (e.g., Vieira et al., 2020), survey methods lack the robustness that "real world" examples of altruism may suggest. For instance, researchers often use kidney donations as a standard for altruism, as it is truly a selfless act Marsh et al., 2014. However, collecting such data can be quite difficult, which is why we propose using tipping behavior as a proxy.

The traditional motivators of tipping behavior are altruism, reward, and duty Ayres et al., 2004; Lynn, 2015. In fact, Jacob et al., 2013 discovered that exposure to altruism prior to tipping significantly increased tipping behavior, evidence to suggest that altruism constitutes a large portion of the drivers of tipping behavior. However, the other two motivators may suggest that tipping behavior is not a completely selfless act as reward and duty are both rational drivers of the behavior. But tipping behavior may be a viable proxy for impure altruism — the idea that altruistic behavior itself is guided by positive feelings from acting benevolently.

2.2 Tipping and COVID-19

Previous research has specifically investigated tipping behavior in response to COVID-19. Lynn, 2021 explores changes in tipping behavior from COVID-19 through two studies: (1) data from a single pizza delivery driver serving a single community in Texas, which may not generalize to a national context, and (2) national data for restaurant transactions that use the credit card platform Square, which is limited only to the food service industry. Additionally, Conlisk, 2022 researches changes in tipping behavior for the transportation industry using data from the Chicago taxi system. Our study explores tipping in the context of gig workers in the delivery industry – rather than food service and transportation industries – at the national level. Duhaime and Woessner, 2019 identify that tipping norms found in other industries do not easily translate over to the gig economy, which is why our study is necessary for understanding this phenomena. Our findings corroborate findings from Lynn, 2021 and Conlisk, 2022 that tipping increased after COVID-19.

3 Methods

3.1 Estimation Strategy

Suppose Y_{it} represents the observed series of interest for units $i = 1, \dots, N$ over time periods $t = 1, ..., T$. Let $Y_{it}(0)$ represent what would have happened if unit i at time t was not treated and $Y_{it}(1)$ represent what would have happened if it was treated. Even though we cannot observe $Y_{it}(0)$ and $Y_{it}(1)$ simultaneously, both are necessary to discover the average treatment effect on the treated units (ATT) in the treatment period, defined as

$$\tau = \mathbb{E}_{it}[Y_{it}(1) - Y_{it}(0) | Y_{it} = Y_{it}(1)]. \tag{1}$$

In traditional causal inference settings, we use data from untreated outcomes after the treatment date, T_0, to estimate $Y_{it}(0)$ Rubin, 2005. However, if *all* N units are treated after T_0, then utilizing untreated units is no longer viable. To overcome this challenge, we utilize pre-treatment observations (i.e., $Y_{it}(0)$ when $t < T_0$) to predict $\hat{Y}_{it}(0)$, which should yield valid counterfactual estimates because pre-treatment values have not been impacted by the treatment. We compare different forecasting techniques that utilize *only* historical observations to identify the best counterfactual predictor. We can then estimate the ATT using $\hat{Y}_{it}(0)$ for $T_0 \leq t \leq T$:

$$\hat{\tau} = \frac{\sum_{i=1}^{N} \sum_{t=T_0}^{T} \left(Y_{it}(1) - \hat{Y}_{it}(0) \right)}{N(T - T_0 + 1)}. \tag{2}$$

3.2 Data Preparation

We used tipping behavior data collected by the digital delivery service Gopuff from May 1, 2019 to May 1, 2020 across the 92 US metropolitan regions that Gopuff is active in. We gathered all transaction data of 50,000 unique customers from 2019 to 2020. All 50,000 customers are considered very active by Gopuff's standards; utilizing customers who became active only after the pandemic would bias the analyses because we could not evidence their pre-pandemic tipping behavior. We also any observations that had no revenue (perhaps because of data collection issues) and observations in which the amount tipped exceeded the amount paid because those are rare and outliers. One important consideration about

Table 1: Associating Tips and Revenue

	Dep. var: Tips ($)
Cost of goods ($)	0.094***
	(0.0002)
Constant	0.155***
	(0.004)
Observations	664,880
R^2	0.361
Adjusted R^2	0.361
Note: HC1 SE	*p<0.1; **p<0.05; ***p<0.01

Gopuff's service is that it does not require drivers or customers to share demographic information, so such details could not be used in our analyses.

On the Gopuff app, aftering confirming purchases, the customer is presented with options to tip the delivery driver $1, $2, $3, or none. Because people are presented with these as default options instead of the traditional tip price in percentage, we tested to see if the amount tipped is associated with revenue, discovering a $0.09 increase in tips for every extra dollar in order cost (p-value < 0.01; see Table 1). Because of this, all subsequent references to "tips" considers tips as a percentage of revenue rather than a raw dollar amount.

In order to predict a trustworthy COVID-period counterfactual, the pre-COVID time series need little missing data. However, in over 90% of the data, customers did not order daily or even weekly (the average number of pre-COVID orders was 13), limiting our ability to use time series methods for counterfactual prediction at the customer level. To circumvent the data paucity problem, we averaged tips by metro region and day. We then filtered out observations in metro regions with less than 15 observations for more than 36 days (10% of the data), as that would suggest an insufficient amount of data to make strong claims about tipping behavior for those days. Afterwards, only eleven metro regions remained (see Figure 1). Because people tend to repeat similar orders on the same day of the week (e.g., people order alcoholic beverages on Friday night every week), missing data was interpolated using the daily average percent tip for that metro region from seven days before.

3.3 Counterfactual Modeling and Inference

We compared three forecasting techniques in order to impute a counterfactual. To establish a baseline, we used the last observation from the training set for each region as the prediction (naive model). We also tested Vector Autoregression (VAR), a statistical technique that captures the relationship between different quantities over time. Considering that each regions' time series is similar, lagged values of tipping behavior from other regions were used as features. We increased the number of lags incrementally until the residuals were uncorrelated with a Portmanteau test Hyndman and Athanasopoulos, 2018. We also considered neural network autoregression (NNAR), a feed forward neural network with one hidden layer that utilizes lagged values as inputs, with a sum of squares objective function. In the

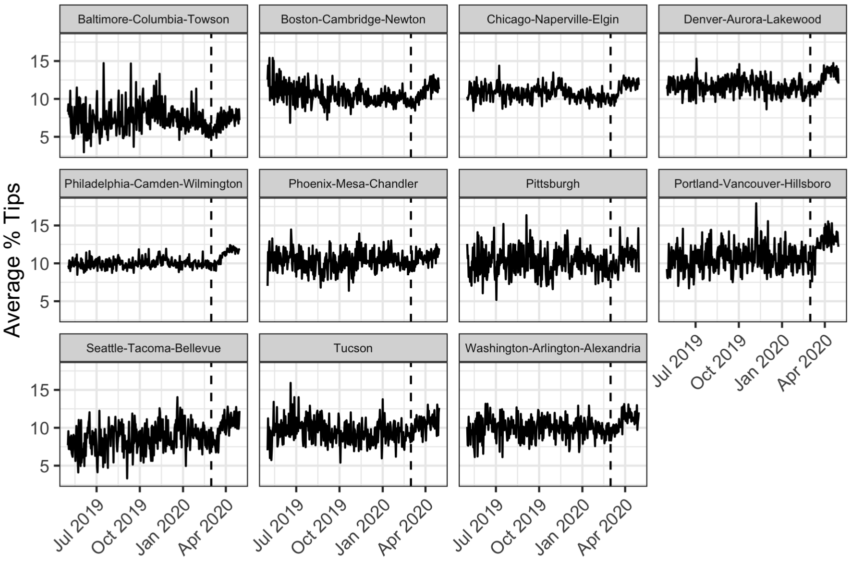

Figure 1: Average daily tips (as a percentage of order cost) per region from May 1, 2019 to May 1, 2020. The dashed line represents the start of COVID-19 – March 1, 2020. After cleaning, only these eleven regions had sufficient data (less than 36 days with less than 15 orders). Even upon visual inspection, it is obvious to recognize a positive shift in average tips after COVID-19 began, especially in the Philadelphia-Camden-Wilmington region.

hidden layer, the number of nodes was equal to the floor of half the number of inputs, and the hidden layer used a sigmoid activation function, which does not fall susceptible to the vanishing gradient problem since there is only one hidden layer. See section A.2 for more details on these methods. Predicting N values of quantity y, denoted as $\hat{y}$, models were then evaluated using mean absolute percent error (MAPE) and root mean squared error (RMSE):

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \text{ RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{3}$$

After all data filtering and aggregation, the dataset was compressed from 600,000 observations to about 4,000 observations – one for each day and metro region. Due to the limited number of observations, we used a cross validation technique made specifically for time-series Hyndman and Athanasopoulos, 2018. In the first iteration, we used observations from May 1, 2019 until October 30, 2019 as the training dataset and observations from October 30, 2019 to December 30, 2019 as the validation dataset. In the second iteration, we added one day to the training set and shifted the validation set over one day; so, we used observations from May 1, 2019 until October 31, 2019 as my training dataset and observations from October 31, 2019 to December 31, 2019 as the validation dataset. We repeated this process until the validation set bled into the COVID-19 period (after March 1, 2020). See section A.1 for more details.

After discovering the best model, we forecast counterfactual time series for the COVID-period of March 1, 2020 to May 1, 2020 for each region. First, we stacked predicted and actual COVID-period values, added an indicator for treatment, and ran the following regression with HC1 standard errors to fix for heteroskedasticity, where β_1 represents the percent impact of COVID-19[1]:

$$log(Tips(\%)) \sim \beta_0 + \beta_1 \mathbb{I}(\text{Observed}) + \varepsilon. \tag{4}$$

4 Results

NNAR had lower MAPE and RMSE scores than the other models (see Table 2). Unsurprisingly, the naive model performed the poorest with a MAPE of 14.35% and RMSE of 0.018. VAR was in-between both. We believe NNAR's performance is due to the regularization term that keeps weights small and the fact that inputs were already between 0 and 1. Both of these help prevent overfitting in neural networks, which is explains why NNAR was able to perform better than VAR.

When considering the impact of COVID-19 on tipping behavior, the regression results show an 8% increase in average daily tips across the eleven regions (p-value < 0.01), suggesting COVID-19 caused an increase in altruistic behavior (see Table 4 and Figure 2). These findings are perfectly in line with that of Lynn, 2015 and Conlisk, 2022, both of which found increases in tipping behavior after the start of the pandemic in various industries. And these results fit in with the theory that altruism is "born of suffering," a belief that hardships motivate people to act kinder towards one another Vollhardt, 2009. Such behavior has also been discovered in times of war, where those exposed to greater violence are more responsive to refugee distress Hartman and Morse, 2015. However, these results contradict survey research that found no change in altruistic behavior caused by COVID-19 at a population level Vieira et al., 2020. That said, Vieira et al., 2020 was aimed at studying the effect of increased threat from COVID-19 (i.e., increase in regional cases) on altruism, which may explain the different conclusions. Perhaps the impact of COVID-19 on altruism was limited only to the presence of COVID-19 itself in regions and not the number of cases.

5 Placebo Test

While the technique used to estimate the causal effect is seemingly intuitive, it is also underutilized because it cannot capture changes due to other interventions. In most policy-related research, there are usually untreated units that can help capture these potential differences; however, this is certainly not the case here. To further prove the validity of this technique, I perform the same set of analyses but with a new 61-day intervention period and a randomly selected start date of October 14, 2019 to see if this study's finds are truly significant or just spurious.

5.1 Placebo Test Methods

Using the cross validation technique outlined in section A.1, we evaluate the same three models. In the first iteration, I used observations from May 1, 2019 until June 15, 2019 as my training dataset and observations from June 15, 2019 to August 8, 2019 as the validation dataset (61 day validation period). In the second iteration, I added one day to the training set and shifted the validation set over one day; so, I used observations from May 1, 2019

[1] A clarifying note: Percent-impact refers to the percent change in tips, which are themselves a percentage of order cost. For example, a 25% impact would suggest an increase in average daily tips from 10% to 12.5% and not a shift from 10% to 35%.

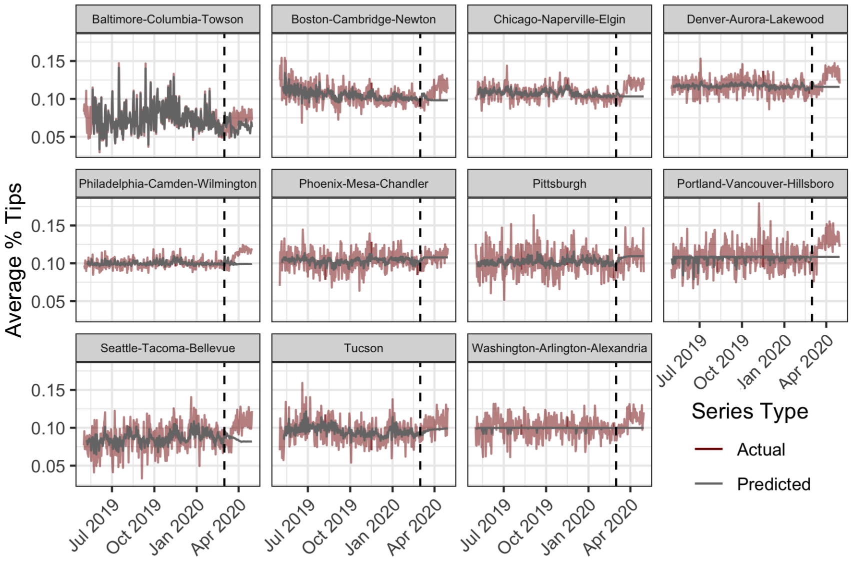

Figure 2: Counterfactual forecasts were developed using NNAR with March 1, 2019 to February 29, 2020 as training set and March 1, 2020 to May 1, 2020 as test set (denoted by dashed line). During COVID-19, average percent tipped *increased* by about 8%, suggesting increased altruism.

until June 16, 2019 for my training set and observations from June 16, 2019 to August 9, 2019 as the validation dataset. I repeated this process until the test set bled into the intervention period starting October 14, 2019. The naive baseline, VAR, and NNAR were compared using MAPE and RMSE again.

After discovering the best model, counterfactuals were predicted for the intervention period for each region using March 1, 2019 to October 14, 2019 as the training period. Causal estimates were derived using the regression outlined in equation 4.

5.2 Placebo Test Results

Again, NNAR had lower MAPE and RMSE scores than the other models (see Table 2), with the naive prediction as the worst model. Additionally, the $\mathbb{I}(Observed)$ coefficient was not significant with a p-value greater than 0.05 (see table 2 and Figure 3), suggesting that there was no treatment effect found in this time period, which makes sense considering there were no national interventions in this time period (to my knowledge). This null result is further evidence for this technique's validity, which should be taken advantage of during COVID-19 to guide policy responses.

6 Discussion

This study discovered that COVID-19 *increased* gig workers' tips in the delivery industry slightly. Importantly, these discoveries can be applied to various policy domains, especially with regards to relief packages and government aid. While the public may have aided gig workers, an average increase of 8% – approximately \$0.32 per order – may not account for the increased hazards associated with working during crises. For example, some large

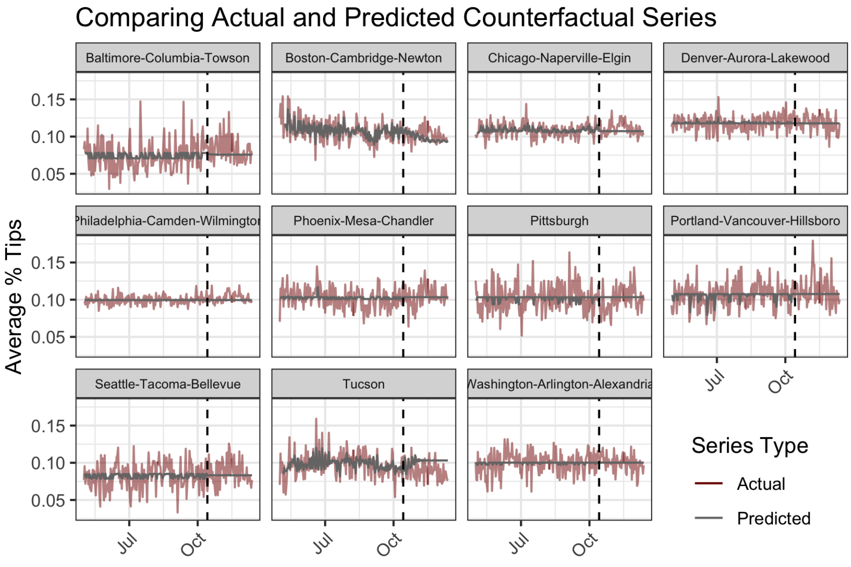

Figure 3: Counterfactual forecasts were developed using NNAR with March 1, 2019 to October 14, 2019 as training set and October 14, 2019 to December 14, 2020 as test set (denoted by dashed line). During the treatment period, average percent tipped did not increase, further validating this technique for estimating causal effects.

companies like Amazon, Kroger, and Whole Foods paid essential workers up to $2 more per hour during the pandemic Kinder, 2022. Drivers from platforms like DoorDash report making an average of 1.5 deliveries per hour, which would not account for the hazard pay. Government aid may be necessary to supplement the increased public support during COVID-19 and in future crises.

These results should be taken cautiously before being extrapolated to the greater gig workforce. Because this study Because this study utilized data from Gopuff (whose target audience is the Millenial generation), these results may not generalize to companies with more diverse customer demographics. Additionally, prosocial behaviors in the delivery industry are quite different than those of other gig industries; for example, tipping is not a standard practice in platforms like Airbnb. Instead, standard practice is to leave gifts or thank you notes to hosts in the rental industry; while the utility or cost of the gifts may increase during disasters, this behavior is not guaranteed. However, because Lynn, 2021 and Conlisk, 2022 found similar results in various industries outside the gig economy, there is evidence to suggest that it is likely that prosocial behavior increased in various components of the gig economy.

One might suggest the increase in average tips was not a result of COVID-19 but rather a reaction to increased activity because businesses were shut down during this period. However, the root cause of these shut downs was still COVID-19, so as long as the root cause of any other intervention was COVID-19, then these findings can still be causal. Because COVID-19 possibly affected individual transactions, comparing transactions before and during COVID-19 may not hold a causal interpretation, which is why we do not explore such analyses. However, the causal interpretations of our findings are still valid.

Future research into the gig economy's resilience in disasters should focus on investigating similar questions in various industries with other proxies for prosocial behavior where

Table 2: Comparing Forecasting Model Performance For COVID-19 and the Placebo Tests

Test	Model	MAPE	RMSE
COVID-19	Naive	14.358	0.018
	VAR	11.258	0.014
	NNAR	10.473	0.013
Placebo	Naive	16.318	0.020
	VAR	13.546	0.017
	NNAR	11.789	0.015

Table 3: Comparing Actual and Predicted Tips (%) During COVID-19 (Left) and During Placebo Intervention (Right)

	Dep. var.: log(Tips (%))	
	COVID-19 Test	Placebo Test
$\mathbb{I}$(Observed)	0.08^{***}	0.002
	(0.001)	(0.008)
Constant	0.155^{***}	-2.310^{***}
	(0.015)	(0.005)
Note:HC1 se	*p<0.1; **p<0.05; ***p<0.01	

tipping is not the norm, such as the rental gig industry. Additionally, understanding whether such increases in prosocial behavior was evidenced during other disasters and conflicts can help guide future policy responses in aiding gig workers. Lastly, research into potential avenues for increasing prosocial behavior towards gig workers will also be useful for marketing campaigns to aid gig workers. Our research marks a strong first step in understanding how disasters moderate prosocial behaviors in the context of the gig economy.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, *105*(490), 493–505.

Alfaro, L., Faia, E., Lamersdorf, N., & Saidi, F. (2020). *Social interactions in pandemics: Fear, altruism, and reciprocity* (tech. rep.). National Bureau of Economic Research.

Ayres, I., Vars, F. E., & Zakariya, N. (2004). To insure prejudice: Racial disparities in taxicab tipping. *Yale LJ*, *114*, 1613.

Bajwa, U., Knorr, L., Di Ruggiero, E., Gastaldo, D., & Zendel, A. (2018). Towards an understanding of workers' experiences in the global gig economy. *Globalization and Health*, *14*(124), 2–4.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, *9*(1), 247–274.

Card, D., & Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

Conlisk, S. (2022). Tipping in crises: Evidence from chicago taxi passengers during covid-19. *Journal Of Economic Psychology*, 102475.

Duhaime, E. P., & Woessner, Z. W. (2019). Explaining the decline of tipping norms in the gig economy. *Journal of Managerial Psychology*.

Girdhar, A., Kapur, H., Kumar, V., Kaur, M., Singh, D., & Damasevicius, R. (2021). Effect of covid-19 outbreak on urban health and environment. *Air Quality, Atmosphere & Health*, *14*(3), 389–397.

Hartman, A. C., & Morse, B. S. (2015). Wartime violence, empathy, and intergroup altruism: Evidence from the ivoirian refugee crisis in liberia. http://cega.%20beerkely.%20edu/assets/miscellaneous%5C_file/119%5C_-%5C_HartmanMorseViolenceEmpathy-May%5C_

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Jacob, C., Guéguen, N., Ardiccioni, R., & Sénémeaud, C. (2013). Exposure to altruism quotes and tipping behavior in a restaurant. *International Journal of Hospitality Management*, *32*, 299–301.

Kinder, M. (2022). Covid-19's essential workers deserve hazard pay. here's why-and how it should work. https://www.brookings.edu/research/covid-19s-essential-workers-deserve-hazard-pay-heres-why-and-how-it-should-work/

Liadze, I., Macchiarelli, C., Mortimer-Lee, P., & Juanino, P. S. (2022). The economic costs of the russia-ukraine conflict. *NIESR Policy Paper*, *32*.

Lynn, M. (2015). Explanations of service gratuities and tipping: Evidence from individual differences in tipping motivations and tendencies. *Journal of Behavioral and Experimental Economics*, *55*, 65–71.

Lynn, M. (2021). Did the covid-19 pandemic dampen americans' tipping for food services? insights from two studies. *Compensation & Benefits Review*, *53*(3), 130–143.

Manyika, J., Lund, S., Bughin, J., Robinson, K., Mischke, J., & Mahajan, D. (2016). *Independent work choice necessity and the gig economy* (tech. rep.). McKinsey Global Institute.

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences*, *111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

Rieger, M. O. et al. (2020). Triggering altruism increases the willingness to get vaccinated against covid-19. *Social Health and Behavior*, *3*(3), 78.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sharif, A., Aloui, C., & Yarovaya, L. (2020). Covid-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the us economy: Fresh evidence from the wavelet-based approach. *International Review of Financial Analysis*, *70*, 101496.

Sundararajan, A. (2017). *The sharing economy: The end of employment and the rise of crowd-based capitalism*. MIT press.

U.s. billion-dollar weather and climate disasters. (2022). https://doi.org/10.25921/stkw-7w73

Vallas, S. P. (2019). Platform capitalism: What's at stake for workers? *New Labor Forum*, *28*(1), 48–59.

Vieira, J., Pierzchajlo, S., Jangard, S., Marsh, A., & Olsson, A. (2020). Perceived threat and acute anxiety predict increased everyday altruism during the covid-19 pandemic. https://doi.org/10.31234/osf.io/n3t5c

Vollhardt, J. R. (2009). Altruism born of suffering and prosocial behavior following adverse life events: A review and conceptualization. *Social Justice Research*, *22*(1), 53–97. https://doi.org/10.1007/s11211-009-0088-1

A Appendix

A.1 Time Series Cross Validation

Suppose we have a time series Y_t with observations for time $t = 1, \ldots, T$. We first use $Y_1, \ldots, Y_{T_1}$ (where $t < T_1 < T$) to predict $Y_{T_1+1}, \ldots, Y_{T_1+k}$. In the next step, we use $Y_1, \ldots, Y_{T_1}, Y_{T_1+1}$ to predict $Y_{T_1+2}, \ldots, Y_{T_1+1+k}$. We repeat this $T - T_1 + k + 1$ times. Figure 4 visualizes the details of this process.

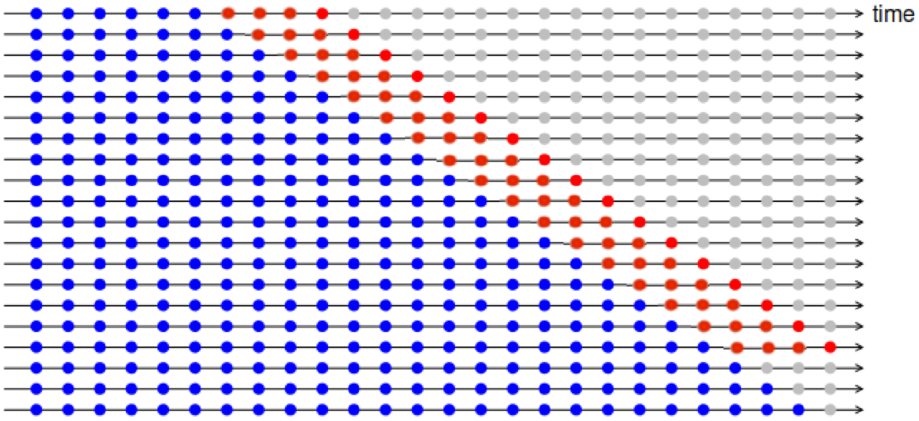

Figure 4: Time series cross-validation steps. The blue points represent the training data while the red points represent a k step forecast. This image was adapted from Hyndman and Athanasopoulos, 2018.

A.2 The Vector Autoregression Model

The vector autoregression model (VAR) is a statistical time series model that allows for the prediction of multiple co-dependent time series.[2] Suppose we have two time series X_t and Y_t with time $t = 1, \ldots, T$. In a simple autoregressive model, we use historical observations of X_t to predict future values using linear regression,

$$X_t = \beta_0 + \gamma_0 X_{t-1} + \ldots + \gamma_k X_{t-k} + \varepsilon, \tag{5}$$

where β_0 is the intercept and ε is the error term. However, suppose X_t and Y_t are correlated with one another. Then, it would make sense to use historical observations of Y_t to predict X_t as well,

$$X_t = \beta_0 + \gamma_0 X_{t-1} + \ldots + \gamma_k X_{t-k} + \eta_0 Y_{t-1} + \ldots + \eta_k Y_{t-l} + \varepsilon, \tag{6}$$

where β_0 and ε are intercept and error terms again. Extending this one step further, if we wanted to forecast Y_t as well, then we could run a multivariate regression,

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{bmatrix} \beta_{00}^1 & \beta_{01}^1 \\ \beta_{10}^1 & \beta_{11}^1 \end{bmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \ldots + \begin{bmatrix} \beta_{00}^k & \beta_{01}^k \\ \beta_{10}^k & \beta_{11}^k \end{bmatrix} \begin{pmatrix} X_{t-k} \\ Y_{t-k} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \tag{7}$$

where the coefficients in the first row of the matrix help predict X_t, the coefficients in the second row of the matrix help predict Y_t, and the superscript on the coefficient represents for

[2] All information on VAR comes from Hyndman and Athanasopoulos, 2018

which lag that coefficient applies to. This is essentially just an extended linear regression. While there are many ways to decide how many lagged terms to use, one technique is to identify the errors are correlated with itself one time period ago (i.e., $Correlation(\varepsilon_t, \varepsilon_{t-1}) \neq 0$), known as serial correlation. Serial correlation suggests that there is some relationship between the data at time t and time $t-k-1$, where k is the number of lagged terms specified in the regression. So adding another lagged term should improve forecast accuracy.

A.2.1 Stationarity

One key assumption of VAR is that time series have a constant mean and variance over time, known as stationarity. We can statistically test for this using the Dickey-Fuller test, whose null hypothesis is that data are non-stationary Hyndman and Athanasopoulos, 2018. As seen in Table 4, the data for each metro region before COVID-19 starts have a p-value less than 0.01, suggesting that we can reject the null and say that data are stationary. So, VAR is a viable option.

Table 4: Dickey-Fuller Test for Stationarity Results For Pre-COVID-19 Time Series

Metro Region	Test Statistic	#Lagged Terms	P-value
Baltimore-Columbia-Towson	4.602	6	< 0.010
Boston-Cambridge-Newton	4.990	6	< 0.010
Chicago-Naperville-Elgin	4.736	6	< 0.010
Denver-Aurora-Lakewood	5.316	6	< 0.010
Philadelphia-Camden-Wilmington	4.377	6	< 0.010
Phoenix-Mesa-Chandler	5.157	6	< 0.010
Pittsburgh	4.878	6	< 0.010
Portland-Vancouver-Hillsboro	6.598	6	< 0.010
Seattle-Tacoma-Bellevue	4.597	6	< 0.010
Tucson	5.293	6	< 0.010
Washington-Arlington-Alexandria	5.240	6	< 0.010

A.3 NNAR

Neural network autoregression (NNAR) is a one-hidden-layer, feed forward neural network specialized for time series models in which we use lagged values of the time series to forecast future observations. We train the neural network the same as we do a normal feed-forward network; however, in the prediction step, predicted values are also used to forecast other values. For example, suppose we have a time series Y_t with $t = 1, \ldots, T$ time periods. Suppose we want to predict values from $T_1 + 1$, such that $t < T_1 < T$, to T (i.e., predict $Y_{T_1+1}, \ldots, Y_T$). Then, we train the NNAR on $Y_1, \ldots, Y_{T_1}$ and update the inputs over time for prediction Hyndman and Athanasopoulos, 2018.

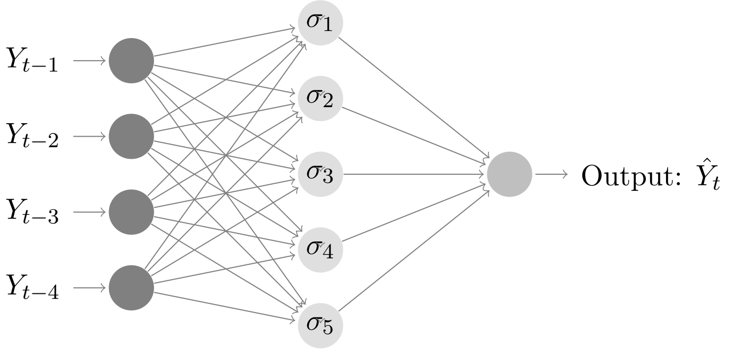

Figure 5: This is an example of NNAR with four-step-lags as inputs and five hidden nodes in the hidden layer. In each hidden node, the inputs, $Y_{t-1}, \ldots, Y_{t-4}$ are linearly combined and passed through a sigmoid activation function. Then, the outputs of the hidden layer are again linearly combined and $\hat{Y}_t$ is predicted. The network udpates using back-propagation, exactly as we do with a multi-layer perceptron. Similar to VAR, NNAR also produces linear outputs, but it could be thought of as a non-linear regression instead.