# The informal economy at times of COVID-19 pandemic

*China Economic Review, 2022*

Feng Guo[a, d], Yiping Huang[b], Jingyi Wang[c] Xue Wang[b, *]

[a] School of Public Economics and Administration, Shanghai University of Finance and Economics, PR China

[b] National School of Development/Institute of Digital Finance, Peking University, PR China

[c] School of Finance, Central University of Finance and Economics; Institute of Digital Finance, Peking University, PR China

[d] Institute of Digital Finance, Peking University, PR China

Reported by Viston Zihao Wang

November 29, 2023

# Outline

# Outline

# 0. Novel issue & perspective

## Novel issue

- The informal economy.

## Novel perspective

- Machine Learning: Gradient Boosting Decision Tree (GBDT).

# Outline

# 1. Introduction

## Background

Tens of millions of **offline micro businesses (OMBs)** have been disproportionately affected by the **COVID-19 pandemic** and **lockdown** measures.

- OMBs operate largely in the informal services, and they are self-employed or informally employed.
- Most OMBs survive with limited savings and lack of access to unemployment benefits.
- Those employed in the gig economy are vulnerable to collapses of income and loss of livelihood.

# 1. Introduction

## Data

Using **weekly** data on around **80 million QR code merchants** from Ant Group.

- **Time span:**
  Dec. 31, 2019 to Apr. 2, 2020 and the corresponding lunar calendar dates in 2018 and 2019.
- **Period defination:**
  The pre- and post-virus periods are the periods before and after Jan. 20 (Dec. 26 in the lunar calendar).
- **Note:**
  (a) Lunar New Year
  (b) Jan. 20, 2020: Human-to-human transmission of the corona virus was confirmed and reported.

西北农林科技大学 NORTHWEST A&F UNIVERSITY

# 1. Introduction

## Predict the counterfactuals using a Machine Learning

- A simple **year-on-year** change in OMB activities $\not\Leftrightarrow$ Real economic
  - The QR code merchants is still on a growing path these two years.
- The **linear DID** specification would lead to a **biased estimation**.
  - It is not clear that
    - the explanatory variables are **linearly** related to OMBs activities?
    - *ex ante* what factors would be **most relevant**?
      Kitchen sink regression $\Rightarrow$ Overfitting & Spurious relationships.
  - Parallel Paths assumption failure.

# 1. Introduction

## Results and conclusions

- OMBs in **urban** areas bore the hardest hit.
- **Female** merchants saw drops.
- **Outsiders** are more vulnerable to shocks.

# 1. Introduction

## Gaps & Contributions

- The **spread**, **containment**, and **economic and political consequences** of COVID-19 and previous pandemics.
- How exposure of Chinese registered firms to the Covid-19 shock varied with a cluster index at the **county level**.

**Contribution #1:** Estimating the **real impact** on hard-hit **informal workers**.

- Informal businesses are inherently difficult to **identify**.
- OMBs **contribute significantly** to employment, especially in developing countries.

**Contribution #2:** **Broadering literature** on informal economy.
**Identifying** informal business by their digital footprints.

# Outline

# 2. Data description

## Description

- **Time span:**
  Dec. 31, 2019 to Apr. 2, 2020 and
  the corresponding lunar calendar dates in 2018 and 2019.
- **Period defination:**
  The pre- and post-virus periods are the periods before and after Jan. 20 (Dec. 26 in the lunar calendar).
- **Note:**
  - Lunar New Year.
  - Jan. 20, 2020: Human-to-human transmission of the coronavirus was confirmed and reported.
  - The 1st week = the 3 days before and 7 days after Lunar New Year's Eve, and the other week = 7 days.

# 2. Data description

## Aggregation

Administrative units, Census tracks, and Other established areas
$\Rightarrow$ Informationin Loss in Big Data.

### Thiessen-polygon

- The method defines an area around a **center point**, where every location is **nearer** to this point than to all the others.
- An essential method for the analysis of **proximity** and **neighborhood**.
- An example on the GeoGebra.org

# 2. Data description

## Aggregation

- **Center point:**
  - **Bank branches** (including self-service branches) within each 500-meter grid cell.
- **Why?**
  - They are densely dispersed in areas with active **businesses** and **economies**.
  - They almost cover **all the places**, except for Shuanghu and Shenzha in Tibet.
- **Result:**
  - Finally establish 138,629 polygons across mainland China

# 2. Data description

## Aggregation

Fig. A.1 shows an example for Chaoyang, Beijing with 437 polygons and red dots.

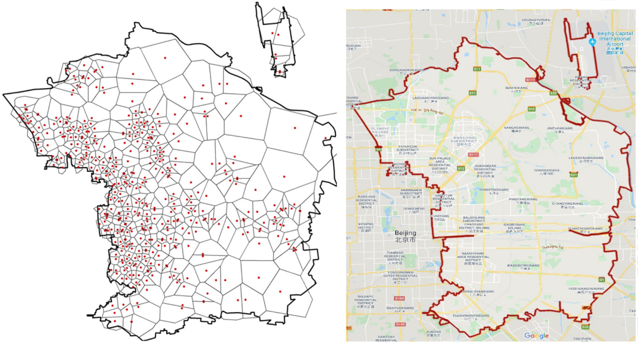



Fig. A.1: Thiessen polygons: Chaoyang District, Beijing.

# 2. Data description

## Raster data

The surroundings of OMBs play a significant role in their daily business.

- **Meteorological conditions:**
  Temperature, Wind speed, Air pressure, Humidity, and Precipitation.
- **Points of Interest (POIs):**
  Hotel, Campsite, Fuel station, Store, or any other specific entity.
- **Cross-section data:**
  Nighttime lights data (500-meter), Population data (1000-meter) and Elevation (30-meter).
- **Transportation convenience:**
  Driving distance from the center of the polygon to that of the County, Prefecture-level city, and Capital city of the province.

# Outline

# 3. Model specification

## Counterfactual OMBs activities

To estimate the **real impact** of the pandemic on OMBs, we need to predict the **counterfactual level** of OMBs economic activities **without the COVID-19 outbreak**.

## Assumption

The activities of OMBs would be stable **if there were no exogenous shock** in a relatively short run.

# 3. Model specification

## Machine Learning: Gradient Boosting Decision Tree (GBDT)

### (a) Training:

$$OMB_{i,2019+k} = F\left(OMB_{i,2018+k}, OMB_{i,2018+(k-1)}, OMB_{i,2018-h}, OMB_{i,2019-h}, X_{i,2019+k}, Z_i\right) \quad (2)$$

where

- $OMB_{i,2019+k}$: the **labeled** number <u>or</u> sales turnover of active OMBs in polygon $i$ in the $k$th week in the post-virus period in 2019 **if there were no pandemic**.
- $OMB_{i,y+k}$: a **vector** of OMB activities (number <u>and</u> sales turnover) in polygon $i$ in the $k$th week in the post-virus period in year $y$.
- $OMB_{i,y-k}$: a **matrix** of OMB activities in polygon $i$ in the three weeks **(i.e., $h = 1, 2, 3$)** before the outbreak in year $y$.
- $X_{i,2019+k}$: a **vector** including the meteorological variables.
- $Z_i$: a **vector** of [POI, Cross-section data, Transportation convenience].

# 3. Model specification

## Machine Learning: Gradient Boosting Decision Tree (GBDT)

### (b) Test:

Table C.2: Performance of the GBDT model.

| | Model: Number of OMBs | | Model: Sales turnover of OMBs | |
|---|---|---|---|---|
| The # week since Jan. 20 | $R^2$ | MAE | $R^2$ | MAE |
| 1 | 0.967 | 11.87 | 0.934 | 0.287 |
| 2 | 0.979 | 11.56 | 0.927 | 0.300 |
| 3 | 0.903 | 13.72 | 0.927 | 0.308 |
| 4 | 0.933 | 12.68 | 0.926 | 0.313 |
| 5 | 0.905 | 13.70 | 0.933 | 0.295 |
| 6 | 0.934 | 12.86 | 0.929 | 0.306 |
| 7 | 0.954 | 12.55 | 0.927 | 0.315 |
| 8 | 0.935 | 13.44 | 0.924 | 0.321 |
| 9 | 0.931 | 14.37 | 0.929 | 0.293 |
| 10 | 0.917 | 13.56 | 0.926 | 0.298 |

# 3. Model specification

## Machine Learning: Gradient Boosting Decision Tree (GBDT)

### (c) prediction:

$$\widehat{OMB_{i,2020+k}} = F\left(OMB_{i,2019+k}, OMB_{i,2019+(k-1)}, OMB_{i,2019-h}, OMB_{i,2020-h}, X_{i,2020+k}, Z_i\right) \quad (1)$$

where

- $\widehat{OMB_{i,2020+k}}$: the **predicted** number or sales turnover of active OMBs in polygon $i$ in the $k$th week in the post-virus period in 2019 **if there were no pandemic**.
- $OMB_{i,y+k}$: a **vector** of OMB activities (number and sales turnover) in polygon $i$ in the $k$th week in the post-virus period in year $y$.
- $OMB_{i,y-k}$: a **matrix** of OMB activities in polygon $i$ in the three weeks **(i.e., $h = 1, 2, 3$)** before the outbreak in year $y$.
- $X_{i,2020+k}$: a **vector** including the meteorological variables.
- $Z_i$: a **vector** of [POI, Cross-section data, Transportation convenience].

# Outline

# 4. Results

## Direct COVID-19 impacts

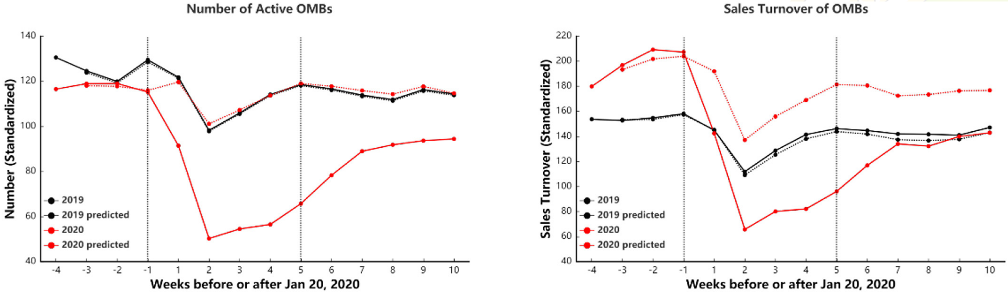



Fig. 1: Actual and predicted OMB activities over time.

Are the predicted counterfactuals **valid**?

# 4. Results

## Direct COVID-19 impacts

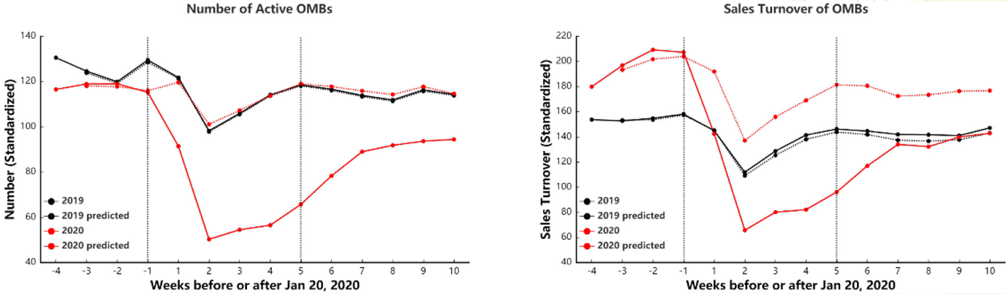



Fig. 1: Actual and predicted OMB activities over time.

**First**, black solid lines vs. black dashed lines.

# 4. Results

## Direct COVID-19 impacts

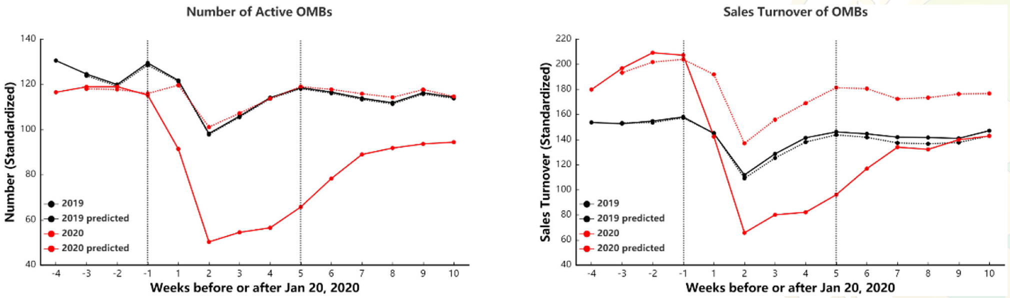



Fig. 1: Actual and predicted OMB activities over time.

**Second**, they assume that the **pseudo-event date** is Dec. 30, 2019 (-3 week).

# 4. Results

## Direct COVID-19 impacts

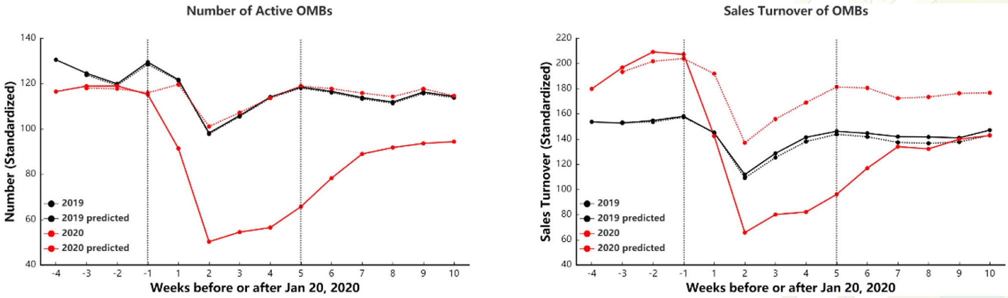



Fig. 1: Actual and predicted OMB activities over time.

**Note #5 of Fig. 1:** The first vertical dashed line marks the first turning point of OMB activities, and the second vertical dashed line marks the start of the bounce of OMBs activities relative to their counterfactuals.

# 4. Results

## Direct COVID-19 impacts

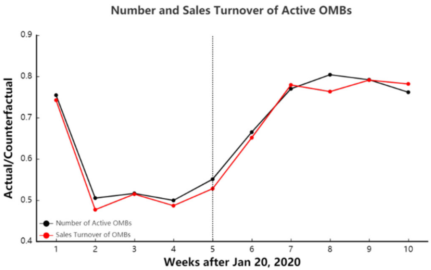



Fig. 2: Changes in OMB activities.

The Ratio of the Actual to the Counterfactual.

- Ratio = 1: No change.
- Ratio was smaller, then the decline was sharper.

Why did the drop be smaller in the first 10 day?

- Lunar New Year. ⇒ Going out of business.
- Stockpiling behavior of consumers.

# 4. Results

## Direct COVID-19 impacts

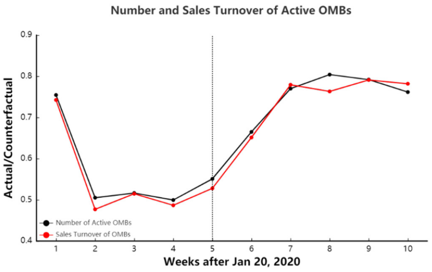



Fig. 2: Changes in OMB activities.

The 2nd week is the worst period:

- the number of OMBs: 50%↓.
- the sales of OMBs: 52%↓.

**Note #4 of Fig. 2:** The vertical dashed line marks the week from which the provincial governments started to revise down the public health emergency response level.

# 4. Results

## Disentangling lockdown effects from the overall impacts

$$OMB_{i,c,k} = \beta_0 + \sum_{j=-3}^{-1} \beta_j Lockdown_{c,j} + \sum_{l=1}^{10} \beta_l Lockdown_{c,l} + \delta X_{i,c,k} + u_c + \nu_k + \varepsilon_{i,c,k} \tag{3}$$

where

- $OMB_{i,c,k}$: the **logarithms** of the number or sales turnove of OMBs in city $c$.
- $Lockdown_{c,j} = 1$ **before** lockdown, and 0 otherwise.
- $Lockdown_{c,l} = 1$ **after** lockdown, and 0 otherwise.
- $X_{i,c,k}$: a **vector** of time-varying control variables, including **newly confirmed cases, new deaths**, and **meteorological variables**.
- $u_c$ and $\nu_k$: the polygon and week **fixed effects**.

# 4. Results

## Disentangling lockdown effects from the overall impacts

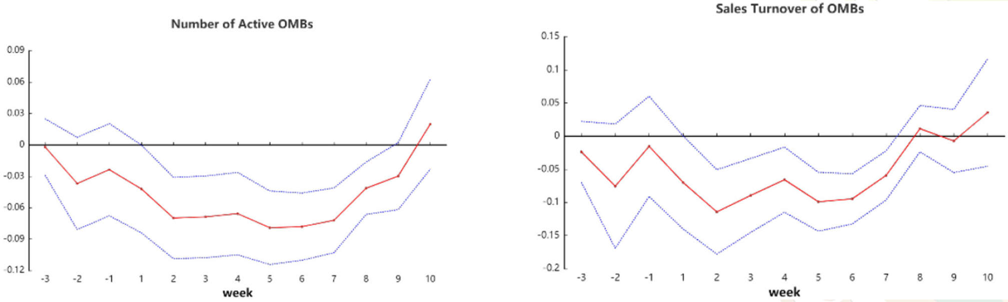



Fig. 3: The dynamic evolution of the lockdown effects.

The estimated coefficients for the lead terms ($j = -1, -2, -3$) are **not significantly different from 0**, then they assume that **the parallel trends across two groups would hold** when there didn't lockdown.

# 4. Results

## Disentangling lockdown effects from the overall impacts

$$OMB_{i,c,k} = \beta_0 + \beta_1 Lockdown_{c,k} + \delta X_{i,c,k} + u_c + \nu_k + \varepsilon_{i,c,k} \tag{4}$$

where

- $OMB_{i,c,k}$: the **logarithms** of the number or sales turnove of OMBs.
- $Lockdown_{c,k} = 1$ **after** lockdown, and 0 otherwise; $k = -3, -2, \cdots, 6$.
- $X_{i,c,t}$: a **vector** of time-varying control variables, including **newly confirmed cases, new deaths**, and **meteorological variables**.
- $u_c$ and $\nu_k$: the polygon and week **fixed effects**.

# 4. Results

Table 1: Impacts of lockdown policies.

| Dependent variables | Log (number of OMBs) | | Log (sales turnover of OMBs) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| lockdown | -0.080*** | -0.059*** | -0.112*** | -0.077*** |
| | (0.016) | (0.012) | (0.026) | (0.021) |
| lagged.case | | 0.000*** | | 0.000* |
| | | (0.000) | | (0.000) |
| lagged.death | | -0.001*** | | -0.001*** |
| | | (0.000) | | (0.000) |
| temperature | -0.008*** | -0.009*** | -0.000 | -0.005 |
| | (0.002) | (0.002) | (0.004) | (0.003) |
| air pressure | 0.006*** | 0.007*** | 0.006* | 0.010*** |
| | (0.002) | (0.002) | (0.003) | (0.003) |
| precipitation | -0.000 | -0.000 | 0.001** | 0.001** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| humidity | 0.001** | 0.001* | 0.002*** | 0.002*** |
| | (0.000) | (0.000) | (0.001) | (0.001) |
| wind speed | 0.001 | -0.003 | 0.008 | 0.009 |
| | (0.005) | (0.005) | (0.012) | (0.011) |
| Thiessen polygon FE | YES | YES | YES | YES |
| Week FE | YES | YES | YES | YES |
| Observations | 1,100,007 | 962,391 | 1,100,007 | 962,391 |
| R-squared | 0.43 | 0.42 | 0.35 | 0.35 |

# 4. Results

## Impacts across urban and rural areas

To match the **granularity** of data, they rely on **nighttime lights** and **population data** to classify the Thiessen polygons into **urban** and **rural** areas.

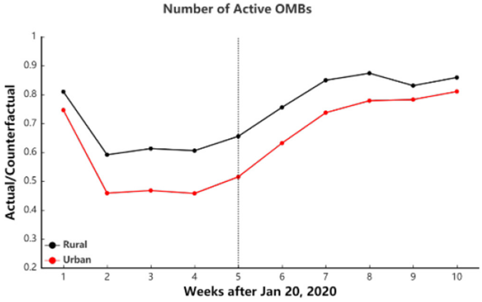
### Method: local-optimized thresholding (LOT)

- Extract nighttime light images for each city with the administrative boundary.
- Use a minimum threshold to segment the images into urban and non-urban areas.
- Record the difference between the extracted area and the reference data (e.g., socioeconomic data, medium- to high-resolution remote sensing data, and so forth).
- By increasing the threshold iteratively, select a threshold that make the difference minimum.

Finally, the **extracted lighting data** is matched with the **population grid data**, and the **OMBs** is divided into urban and rural groups.

# 4. Results

## Impacts across urban and rural areas



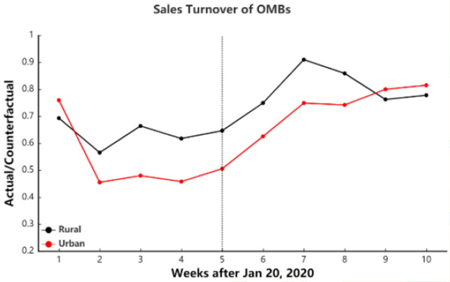




Fig. 4: Changes in OMB activities: urban versus rural.

Fig. 4 shows that **urban areas were hit harder**, compared to OMBs in rural areas.

# 4. Results

## Impacts by owners demographics: Genders

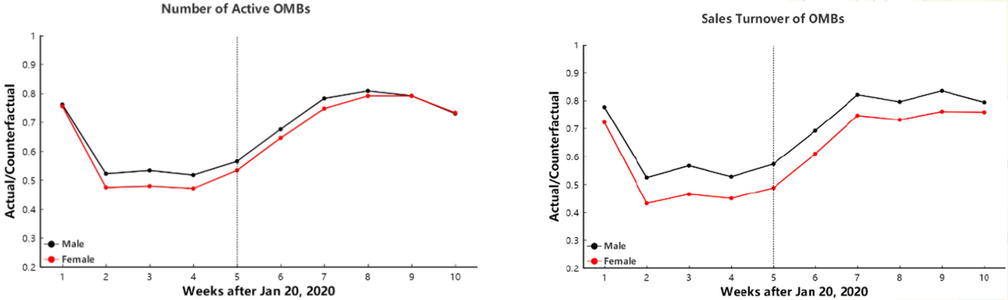



Fig. 5: Changes in OMB activities: male versus female.

Fig. 5 shows that the **female business owners experienced a sharper decline**.

# 4. Results

## Impacts by owners demographics: Place of birth

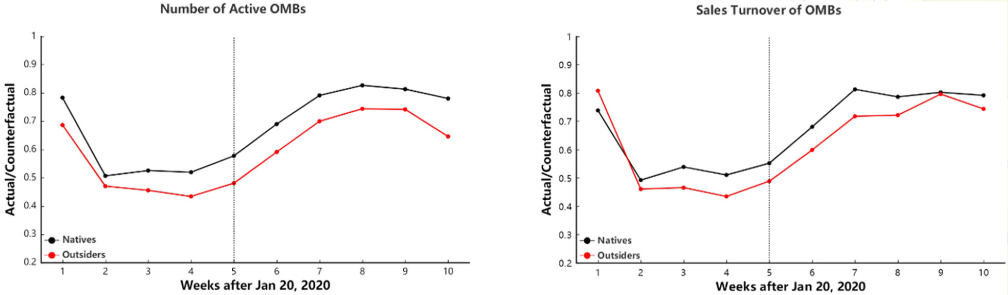



Fig. 6: Changes in OMB activities: outsiders versus natives.

Fig. 6 illustrates the decrease in economic activities were **larger for the outsiders**.

# Outline

# 5. Conclusions and implications

## Conclusions

- The **activities of OMBs** experienced immediate and dramatic collapse, with the biggest weekly contraction of around **50%**.
- The decline due to **lockdown policy** was modest and negligible.
- OMBs in **urban areas** experienced a sharper contraction during the trough.
- **Female** merchants were hit harder than the male merchants.
- The **most vulnerable workers in the gig economy** were hit hard by the COVID-19 pandemic.

# 5. Conclusions and implications

## Implications

- The **quick recovery** of OMBs since the nationwide encouragement of work resumption provides evidence of **the necessity of prioritizing containment** of the virus and **the importance of government support in reopening the economy**.
- They suggest a more continuous policy response to **ensure adequate support for the most vulnerable** at a relatively longer-term amid the new normal of epidemic prevention and control.

# Thank you for your attention!

Reported by Viston Zihao Wang

November 15, 2023