EFFECT ASSESSMENT OF LARGE-SCALE EVENTS VIA SPATIOTEMPORAL APPROACH

Xinwei Chai[1,*], Xian Guo[2], Jihua Xiao[1], Jie Jiang[2]

[1] Beijing Unistrong Science & Technology Co., Ltd, 100176 Beijing, China - (xw.chai, jh.xiao)@chinalbs.org
[2] Beijing University of Civil Engineering and Architecture, 102616 Beijing, China – (guoxian, jiangjie)@bucea.edu.cn

Commission III, WG III/1

KEY WORDS: Effect Assessment, Co-location Analysis, Difference-in-Differences (DID), Spatiotemporal Big-data.

ABSTRACT:

Together with rapid development of location-based services and big-data platforms especially in urban areas, huge amount of spatiotemporal data are collected without properly used; on the other hand, state-of-the-art quantitative policy effect assessment techniques usually require panel data as input. To solve both issues, this paper follows the following approach: obtaining panel data by aggregating spatiotemporal data and feeding them to the effect assessment module. With the help of high-performance computing techniques which are able to deal with huge amount of data, we build framework Aggr-analysis which applies clustering algorithms to shrink the raw data set and find associations between different data sets *via* co-location analysis. Finally, we prove the effectiveness by an example: analysis of resident activities during the COVID-19 Pandemic. We apply Aggr-analysis to process the share-bike usage data and POI (Point Of Interest) data in Beijing, then obtain the panel data required by DID (Difference-in-Differences) method. Supplemented with environmental data, we conclude the net effect of the COVID-19 breakout on society and economy - the pandemic has reduced the overall resident mobility by 64.8% within two months.

1. INTRODUCTION

Studies based on DID design (Difference-in-Differences) are prevalent in the domain of econometrics and quantitative researches (Baker et al., 2022). Due to possible endogeneity problems, part of DID configurations are not valid (Bertrand et al., 2004). The endogeneity concern derives from non-randomness, such as laws or interventions themselves aiming at influencing current situation. Randomization is often infeasible in social scientific researches due to logistical or ethical concerns and so studies rely on observational data.

However, non-periodical emergencies allows for "legal" randomized experiments, as the occurrence can be considered irrelevant to control variables. Studies of such emergency impact usually rely on specific spatiotemporal data (Goodchild and Glennon, 2010, Horanont et al., 2013, Huang et al., 2015, Yu et al., 2018), especially during the ongoing COVID-19 pandemic. To evaluate the effect on the spatiotemporal changes, state-of-the-art approaches use LBS (Location-based services) data as an important data source, such as:

- Social media: researches on Weibo (main microblog social media in China) (Yin et al., 2020, Zhao et al., 2020) explore public attention and information propagation on social networks; Data for Good project organized by Facebook is proved to carry additional location information which is helpful to evaluate the risk of future COVID-19 outbreaks (Chang et al., 2021). However, geo-tagged posts comprise a small part of the whole data, and do not reflect people's routines.
- Mobile phone data-related studies (Zhou et al., 2020, Yabe et al., 2020, Xiong et al., 2020) confirm the positive relationship between human mobility and COVID-19 infections, but one can scarcely distinguish purposive movements from random wandering/indoor movements in these data.
- Navigation data from mapping platforms: studies based on Google Maps (Li et al., 2021) and Baidu Maps (Huang et al., 2020) reveal transportation-related behaviors with navigation data. For the same reason as in social media, navigation data cannot cover regular short movements.

The formerly mentioned studies confirm certain assertions but not valid enough, because their data sets cause endogeneity problem due to very limited coverage of the studied population. In this paper, the authors propose a general framework Aggr-analysis for effect assessment with location-based data. As is often commented, "garbage in, garbage out", Aggr-analysis takes data with unbiased coverage of the research target as input, then converts the input to staggered panel data and use DID technique to analyze the net effect. The evaluation of Aggr-analysis is carried out by a the analysis of impacts of COVID-19 on Beijing based on long-time-sequenced shared-bike data. The rest of this paper is organized as follows. Section 2 details Aggr-analysis. Section 3 reports the spatiotemporal characteristics of the shared-bike usage and reveals the DID results. Section 4 concludes with certain remarks.

2. METHODOLOGY

To make use of big spatiotemporal data in effect assessment techniques, we propose a generalized framework Aggr-analysis (Aggregation-based analysis), which covers the whole process from handling raw input data to reaching final conclusion.

2.1 Workflow

The overall workflow is shown in Figure 1. We first feed Aggr-analysis with input data sets composed of two parts:

1. After clustering, spatiotemporal data are converted to aggregated time-series, serving as explanatory variables.

*Corresponding author

2. Supporting data serve as control variables.

In Analysis module, after denoising, the aggregated time-series are transfered to co-location analysis and transformed to aggregated staggered panel data, which is acceptable for DID model. This aggregated data can also be visualized *via* GIS tools for a better global understanding. DID model takes panel data and control variables as input, resulting in the targeted net effect. Additional with the co-location patterns obtained by co-location mining techniques, we reach a conclusion.

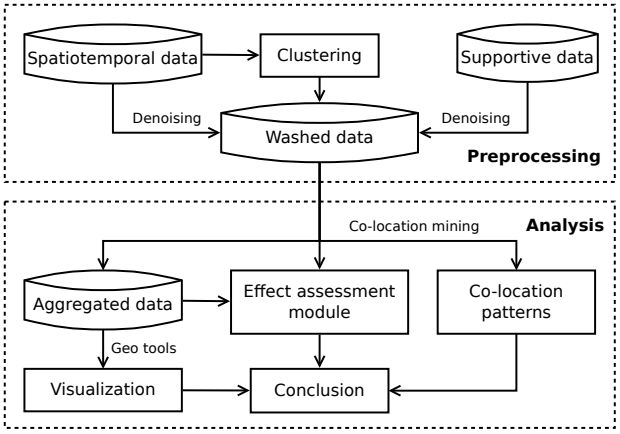


Figure 1. Workflow of Aggr-analysis

2.2 Co-location mining

Co-location patterns hidden in spatial data are collected by co-location mining (Yoo and Bow, 2012). They are usually in form of cliques, as they depend on the definition of neighborhood to describe the mutual relationship of spatial objects and the prevalence threshold. We formally define co-location pattern and its related terms.

Definition 1. *Let $E = \{e_1, \ldots, e_N\}$ be the set of events, classified in categories $C = \{c_1, \ldots, c_n\}$: $\forall e_i$, $\exists c_j$ s.t. $e_i \in c_j$ and $\nexists c_k$, $k \neq j$ s.t. $e_i \in c_k$. A co-location is $X = \{e_1, \ldots, e_k\} \subseteq E$ and a co-location pattern is the smallest category set $P \subseteq C$: $\forall X, X \subseteq \bigcup_{c \in P}$ and $\nexists P' \subseteq P$.*

Definition 2. *Given co-location pattern P and its corresponding co-location instances $\{X\}$, the participation ratio PR is the fraction of events in c_i participating P: $PR(P, c_i) = |\{e_j | e_j \in c_i \wedge e_j \in \cup\{X\}\}|/|\{c_i\}|$, based on which the participation index PI is defined as $PI(P) = \arg\min_{e_i \in P} PR(P, e_i)$.*

A co-location pattern might cover the whole set of events or appears only once which is hardly called a "pattern". One restricts the patterns by setting a threshold of participation index, suggesting that a co-location pattern P appears at least with probability $PI(P)$.

2.3 Generalized DID

In Effect assessment module, one often applies generalized DID model defined as follows:

Let Y_{ist} be the outcome of interest for individual i in group s at time t and T_{st} be a dummy taking value 0 or 1 for whether the intervention has affected group s at time t. One then estimates the following regression using OLS (Ordinary Least Squares):

$$Y_{ist} = \alpha + cX_{ist} + \beta T_{st} + \epsilon_{ist} \tag{1}$$

where $\alpha = A_s + B_t$, A_s and B_t are fixed effects for the groups and years and X_{ist} represents the relevant individual control variables. The estimated impact of the intervention is then the OLS estimate $\hat{\beta}$. Standard errors ϵ_{ist} around $\hat{\beta}$ are OLS standard errors after accounting for the correlation of shocks within each state-year cell.

Y_{ist} is set to the logarithm of the target effect ($\log Y_{ist}$), reducing the absolute error due to singular values and more importantly, regression coefficient β becomes the ratio between the changes of response and explanatory variables:

$$Y_{ist} = \exp(\alpha + cX_{ist} + \epsilon_{ist}) \cdot \exp(\beta) \tag{2}$$

Let $\exp(\alpha + cX_{ist} + \epsilon_{ist}) = C$, given $T_{st} = 1$, we have $Y_{ist} = C \cdot \exp(\beta_2)$, where C is in fact the expected value of Y_{ist} without intervention. The ratio of change due to intervention is:

$$\frac{C - Y_{ist}}{C} = 1 - \exp(\beta_2) \tag{3}$$

2.4 K-segmentation

A study period is divided into several stages according to intuition, for example, pre-intervention and post-intervention of pandemic outbreak. However, the time point when certain intervention take place and when it take effect are probably different, and remain challenging.

In machine learning tasks, one often aims at minimizing the predefined loss function to formulate the best classification such that the elements within the same cluster are similar and the elements across clusters are different (Vladimir, 2002). Likewise, we tried to segment the study period into phases comprising the most similar patterns using K-segmentation.

Definition 3 (K-segmentation)**.** *Let $X = \{x_1, x_2, \ldots, x_N\}$ be a time series of length N. Given $k \in \mathbb{N}$, $k < N$ and index set $\mathbf{T} = \{n_0, \ldots, n_k\}$ with $n_0 = 0$, $n_k = N$ and $\forall i$, $n_i < n_{i+1}$, a K-segmentation of X is the set of time series $X_i = \{x_{n_i+1}, \ldots, x_{n_{i+1}}\}$ where $0 \leq i \leq k-1$.*

To evaluate a K-segmentation, we use $\sigma = \sum_{i=1}^{k} \sigma_i$ as the loss function where σ_i is the standard deviation of division X_i. The goal is to find the best $\mathbf{T}$ to minimize σ, *i.e.,* $\arg\min_{\mathbf{T}} \sigma(\mathbf{T})$. This problem can be solved at the complexity level of $O(N^2 k)$ (Terzi and Tsaparas, 2006). In case k and N are small, the optimum can be found *via* exhaustive search.

3. APPLICATION

A typical application of Aggr-analysis is exploiting the spatiotemporal changes in the human mobility under the influence of COVID-19 pandemic.

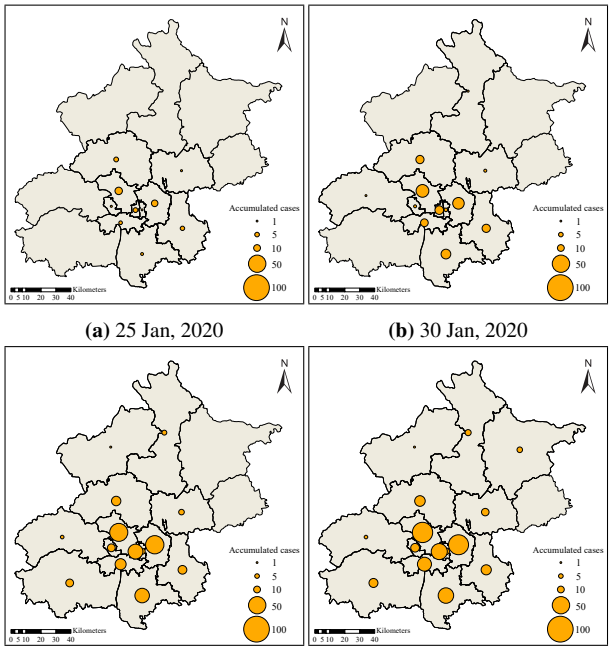
Considering the need of wide and unbiased coverage of target population, wide-spread bike sharing system (BSS) in China is a valid data source for analyzing human mobility *at city-scale* during the pandemic period, as shared-bikes are already an alternative for fulfilling people's need for regular short-distance transportation.

After the COVID-19 outbreak at the beginning of year 2020, social distancing and home quarantine were imposed for the prevention of the pandemic. Additionally, there was a suspension of buses and taxis for a short period. These strict control measures inevitably narrowed the options of public transit, increasing the use of shared-bikes.

3.1 Data set

3.1.1 BSS records This OD (Origin-Destination) data set comes from 1.02 million shared-bikes provided by 4 main BSS operators (Mobike, DiDi Bike, Hellobike, and Ofo) in Beijing. Its records date from Mar, 2019 to Mar, 2020 (66.8 GB) and cover 1.5 million uses per day contributed by 11 million users, which account for half of the total population of Beijing. They are created when users locked/unlocked their shared-bikes, excluding those of rebalancing operations. This exclusion guarantees that the records were collected purely from users. In addition, this data is anonymous and does not cause privacy concerns. It should be noted that the BSS records in some districts (Chaoyang, Fengtai, and Shijingshan) are not available due to the local policies.

3.1.2 Points Of Interest (POIs) POIs of Beijing are collected from APIs provided by AutoNavi[1], a web-mapping platform in China. Each entry has the following attributes: a unique ID `object_id`, an address including lat/lon information, and a three-level classification: `large_category`, `mid_category`, and `sub_category`. Among these categories, we choose seven mid ones: residential area (RA), high-tech company (HC), other company (OC), subway station (SS), shopping plaza (SP), supermarket (SM), and tertiary hospitals (TH)[2].



(a) 25 Jan, 2020 **(b)** 30 Jan, 2020

(c) 05 Feb, 2020 **(d)** 10 Feb, 2020

Figure 2. Snapshots of accumulated confirmed cases in Beijing from 25 Jan to 10 Feb, 2020.

3.1.3 Confirmed COVID-19 Cases In order to exploit the spatial effect of COVID cases in the human mobility nearby, we also collect the confirmed COVID-19 cases during the outbreak from the daily update on the COVID-19 outbreak dashboard provided by the Foreign Affairs Office of the People's Government of Beijing Municipality[3]. The cumulative daily counts of clinically diagnosed cases in each district from 20 Jan to 05 Mar, 2020, and we obtain a total of 87 infected residential areas. According to the timeline of the outbreak (Li et al., 2020), the evolution of the overall pandemic situation of Beijing is shown in Figure 2.

3.1.4 Weather data The weather conditions may also effect human mobility. Herein, the weather data are obtained from the China Meteorological Data Service Center[4], providing daily weather information such as temperature, wind speed, and precipitation, covering the same dates as aforementioned three data sets.

3.2 Tools

Parallel computing performs spatial queries on the big data set:

1. HDFS (Hadoop Distributed File System)[5]: a distributed file system which is suitable for parallel computing.
2. Spark[6]: a parallel analytic engine for big data, and can invoke Structured Query Language (SQL) to process temporal queries in the BSS data, performing denoising and statistical analysis.
3. Apache Sedona[7], formerly GeoSpark (Huang et al., 2017): a GIS-based engine based on Spark, capable of performing spatial analysis and visualization of geo-based data.

We implement our approach in Python and Scala, with certain code available[8]. Cartographic visualization is done by ESRI ArcGIS[9]. All computation is run on a server consisting of 7 machines with Intel®Xeon®, CPU E5-2640 v2 @2.00 GHz, 8 cores, 61.7 GB RAM, 20 MB cache.

3.3 Method

We first cluster certain types of POIs using DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) since these POIs may be located close to each other (*e.g.* RA and SM) causing duplicate patterns in further analysis. Together with BSS data, POI data are to play the role of response variable reflecting human mobility in various aspects, while weather data act as control variable. Confirmed cases with time are to be used in the calibration of the study period. While in Analysis module, there are five tasks:

1. Statistical charts: a first-step visualization of the statistics of the shared-bike data;
2. Co-location analysis: aggregating shared-bike usage near different POIs (within range of 100 m) in different phases, generating panel data for DID analysis;
3. Phase segmentation: using K-segmentation approach to divide the entire study period into logical phases by minimizing the loss function;
4. DID: quantitative analysis of the impact of COVID-19 reflected by shared-bike usage change;
5. Heatmap: visualization of the shared-bike data based on the aspect of space & time.

By synthesizing the above results, we reveal the net human mobility change during the pandemic. Among the steps above, DID configuration should be introduced in more details. As the COVID-19 pandemic outbreaks during the Chinese New Year

[1] `https://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/WSXCNM`
[2] Tertiary hospitals are considered as the top-class hospitals in China.
[3] `http://wb.beijing.gov.cn/home/ztzl/kjyq/`
[4] `http://data.cma.cn/en`
[5] `https://hadoop.apache.org`
[6] `https://spark.apache.org`
[7] `https://sedona.apache.org`
[8] `https://github.com/XinweiChai/bike_analysis`
[9] `https://www.esri.com/en-us/arcgis`

of 2020, by taking the shared-bike usage of 2019 as a control variable (a "virtual pandemic" during the Chinese New Year of 2019), we configure the DID model as shown in Table 1.

$T \times D$	Before pandemic ($D = 0$)	During pandemic ($D = 1$)
2019 ($T = 0$)	0	0
2020 ($T = 1$)	0	1

Table 1. DID configuration.

We construct the DID regression model Equation 4 by concretizing Equation 1:

$$\log U_t = \alpha + \beta_1 \cdot before_{2020,t} + \beta_2 \cdot during_{2020,t} + \theta_t + \epsilon_t \quad (4)$$

where t is the date and U_t is the shared-bike usage on day t. $before_{2020,t}$ and $during_{2020,t}$ are dummy Boolean variables: $before_{2020,t}$ takes 1 if t is 4 to 11 days before the outbreak of pandemic. This term is set to verify the common trend assumption in DID analysis. $during_{2020,t}$ takes 1 when t is during the pandemic or in mitigation period (corresponding to $T \times D$ in Table 1). α is a constant term, β_1, β_2 are fitted coefficients, θ_t is the date fixed effects (weather, temperature, weekday/weekend, Chinese New Year, *etc.*), and ϵ_t is the residual term. The effect of holiday and pandemic is evaluated by β_1 and β_2.
According to Equation 2 and Equation 3, during the pandemic, as $before_{2020,t} = 0$, $during_{2020,t} = 1$, let $\exp(\alpha+\theta_t+\epsilon_t) = C$, the proportion of human mobility reduction reflected by shared-bike usage done by pandemic effect is:

$$\frac{C - U}{C} = 1 - \exp(\beta_2) \quad (5)$$

3.4 Result

In this section, we verify the effectiveness of Aggr-analysis *via* assessing the effect of large-scale events through the COVID-19 outbreak in Beijing. Specifically, statistical analysis and visualization are firstly employed to present the data characteristics, followed by discussions in k-segmentation and DID-analysis.

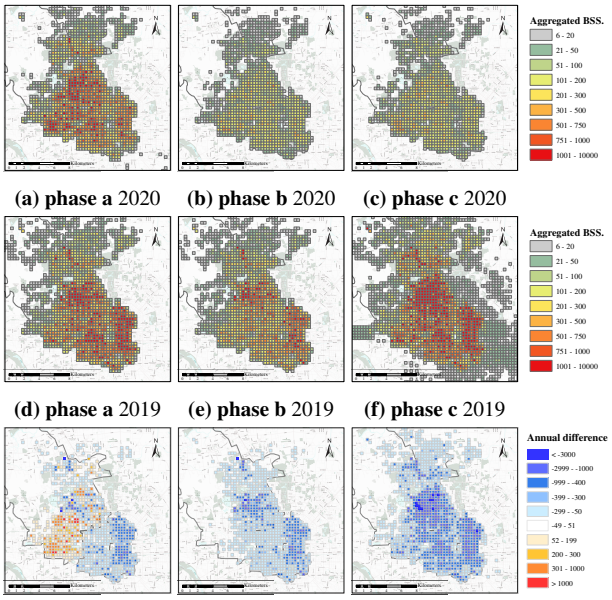
3.4.1 Statistical Analysis Table 2 presents in a statistical aspect of aggregated bike usages before/after the COVID-19 outbreak. The upper part shows that during rush hours of working days, the shared-bike usage in 2020 is of the same scale as that of 2019, suggesting that shared-bike use demand follows common trends which is required by DID. However, the lower part depicts the case of Chinese New Year holiday, where the overall shared-bike usage drops to less than 40% compared to the same period in 2019, suggesting more companies stops working due to Chinese New Year holiday in 2020.

3.4.2 Visualization Figure 3 delineates the sum of aggregated shared-bike usage in **phase a, b, c** of 2020 (row 1), the shared-bike usage in the corresponding period of 2019 (row 2), and the difference of the former two results (row 3).
Figure 3d summarizes the average shared-bike usage intensity during the same time interval as **phase a** in 2019, showing similar spatial patterns as in 2020. Figure 3g shows the pre-pandemic difference of bike usage between 2019 and 2020, which is irrelevant to the COVID-19 outbreak.

3.4.3 K-segmentation According to Definition 3, we identify key time points using the shared-bike usage data. We apply elbow method and determine the best-classified phases at $k = 3$

	Daily average shared-bike usage ($\times 10^5$)		
Phase	08:00 - 09:00 (weekdays)	18:00 - 19:00 (weekdays)	All-day
02 Jan - 20 Jan 2019	2.46 ± 0.59	1.30 ± 0.39	15.0 ± 4.5
02 Jan - 20 Jan 2020	2.58 ± 1.10	1.37 ± 0.74	12.7 ± 6.3
Phase	08:00 - 09:00 (whole week)	18:00 - 19:00 (whole week)	All-day
04 Feb - 10 Feb 2019	0.30 ± 0.05	0.24 ± 0.08	4.90 ± 1.56
24 Jan - 02 Feb 2020	0.10 ± 0.02	0.12 ± 0.03	1.72 ± 0.35

Table 2. Shared-bike usage in different time intervals, shown in form $\bar{x} \pm 2\sigma$.



(a) phase a 2020 **(b) phase b** 2020 **(c) phase c** 2020
(d) phase a 2019 **(e) phase b** 2019 **(f) phase c** 2019
(g) Diff. **a** and **d** **(h)** Diff. **b** and **e** **(i)** Diff. **c** and **f**

Figure 3. Comparison of shared-bike usage between 2019 and 2020 in different phases.

given the study period of 62 days. The three consecutive phases are defined *via* the split points of 23 Jan and 24 Feb, and they can be named intuitively: **phase a** before the pandemic, **phase b** during pandemic, **phase c** pandemic mitigation.
We verify the segmentation by applying K-segmentation respectively on different categories of POI, as shown in Table 3. The segmentation on sub-categories is consistent with that of overall share-bike usage. There is a minor difference between HC, OC and that of other categories. This difference is explained by the lag between the urgent shutdown and the start of the Chinese New Year vacation. Social and productive activities have not resumed until 24 Feb, which is already two weeks after the official declaration of the partial restart.

3.4.4 DID Analysis Table 4 presents the OLS regression result of Equation 1.
We do not list the constant term α since it is the intercept and we focus more on the *change* of the shared-bike usage which reflects the human mobility of residents. $|\beta_1|$ is small enough, suggesting that the effects of the Chinese New Year and that

Category	HC	OC	RA	SS
Split point 1	23.01	23.01	24.01	24.01
Split point 2	24.02	28.02	24.02	24.02
Category	SP	SM	TH	Overall
Split point 1	24.01	24.01	24.01	23.01
Split point 2	24.02	24.02	24.02	24.02

Table 3. Period segmentation of 02 Jan to 02 Mar, 2020

	Overall	RA	HC	OC
β_1	0.033 (0.066)	0.011 (0.069)	0.051 (0.114)	0.023 (0.1)
β_2	-1.044 (0.125)	-0.889 (0.136)	-1.355 (0.214)	-1.183 (0.189)
R^2	0.921	0.906	0.894	0.892
$1-e^{\beta_2}$	64.80%	58.89%	74.21%	69.36%
	SS	SP	SM	TH
β_1	0.027 (0.084)	0.005 (0.166)	-0.01 (0.083)	0.017 (0.067)
β_2	-1.51 (0.156)	-1.331 (0.249)	-0.985 (0.143)	-0.886 (0.165)
R^2	0.911	0.687	0.824	0.916
$1-e^{\beta_2}$	77.91%	73.58%	62.66%	58.77%

Table 4. The net effect of COVID-19 on shared-bike usage obtained *via* DID, with all the value of $During_{2020,t}(\beta_2)$ at 5% significance level, shown in form $\bar{x}(\sigma)$

of the shared-bike usage trend are well absorbed by the date fixed effects θ. β_2 is negative and statistically significant, suggesting that the pandemic reduces human mobility compared to year 2019. We estimate the proportion of shared-bike usage decrease due to the net impact of the pandemic *via* $1-\exp(\beta_2)$. β_2 for all the POI categories is -1.044 implying that the pandemic reduces the overall bike usage by 64.8%. This percentage is slightly lower than 69.49% estimated for Wuhan (the ground zero), as reported in (Fang et al., 2020). (Mu et al., 2020) estimates the intra-city mobility reduction in Beijing to be between 56.5% and 65.2%, which is consistent with our result (64.8%). Analogous analyses are conducted respectively in the chosen POI categories. With all β_2 values at a 5% significance level, we are prone to believe that COVID-19 pandemic has a negative impact on human mobility close/belonging to urban core functional areas. The principle cause is attributed to the municipal restrictions especially wide-spread quarantine. The estimated mobility reduction *due to COVID-19 pandemic* of SS (77.91%), HC (74.21%), and SP (73.58%) are higher than that of other categories.

3.4.5 One Step beyond DID As certain POI categories often co-locate with each other, generating co-location patterns, *e.g.*, there are usually convenience stores downstairs an office building. To distinguish the motivation of each shared-bike usage, we classify different purposes by characteristic timeslots, *i.e.*. For instance, we selected 8:00 - 10:00 as the characteristic timeslot for HC and OC.

Table 5 shows the shared-bike usage per day within range of 100 m of POIs during its characteristic timeslot, where RA, HC, OC are cluster-based (preprocessed *via* DBSCAN) and SS, SP, SM, TH are point-based. U_a, U_b, U_c stand for the shared-bike usage during **phase a, b, c**. Two associated ratios measure the extent to which the COVID-19 epidemic has led: U_b/U_a quantifies the decline in shared-bike usage during the Chinese New Year, which coincides with the strict quarantine period, U_c/U_a reflects the recovery progress afterwards (% of pre-pandemic). We reach some interesting implications, for example:

POI category	Time slot	# POIs (# clusters)	Bike usage per POI($\times 10^3$)			Ratio	
			U_a	U_b	U_c	U_b/U_a	U_c/U_a
RA	all-day	5657(1204)	254.4 ± 133.8	61.65 ± 35.18	99.30 ± 5.59	24.2%	39.0%
HC	8-10h	3858(81)	4.06 ± 3.52	0.42 ± 0.56	1.16 ± 0.15	10.4%	28.6%
OC	8-10h	32301(577)	32.69 ± 26.68	4.19 ± 4.54	9.95 ± 1.11	12.8%	30.4%
SS	8-22h	137	34.07 ± 21.57	4.25 ± 3.15	8.33 ± 0.61	12.5%	24.5%
SP	18-20h	217	3.70 ± 2.01	0.79 ± 0.62	1.38 ± 0.15	21.3%	37.2%
SM	18-20h	1076	13.59 ± 7.63	3.43 ± 2.44	6.27 ± 0.58	25.2%	46.1%
TH	all-day	86	13.69 ± 7.14	3.38 ± 1.79	5.00 ± 0.19	24.7%	36.5%
Overall	all-day	———	547.3 ± 301.7	124.0 ± 66.3	195.8 ± 10.4	22.7%	35.8%

Table 5. Bike usage in **phase a, b, c** around chosen POIs, shown in form $\bar{x} \pm 2\sigma$ (see abbreviations in Appendix).

- Among the values of U_b/U_a, HC has the biggest decrease (down to 10.4%), corresponding to the fact that workers in HC agree most with "work from home";
- As for U_c/U_a, 28.6% for HC and 30.4% for OC suggest that the recovery from the peak of panic is not satisfying;
- SM has the highest recovery ($U_c/U_a - U_b/U_a$), showing the people's basic needs are of extreme importance even in strict quarantine period.

4. CONCLUSION

This paper introduces our framework Aggr-analysis designed for effect assessment of large-scale events from spatiotemporal perspective. Fed with huge amount of LBS data, it carries co-location analysis and produces panel data with the help of parallel computing, serving for quantitative researches. DID, a representative statistical technique, consumes the panel data and finally obtains the *net* impact of the targeted object.

An application of Aggr-analysis is conducted on the impact assessment of COVID-19 pandemic in Beijing after the outbreak. We conclude an overall decrease of human mobility to be 64.8% and the impact of COVID-19 pandemic lasts at least till the end of our study period.

Given non-biased and full-coverage raw data, Aggr-analysis is a *generalized* tool for effect assessment, also applicable in econometric researches and policy making. In the ongoing research, we are extending Aggr-analysis to predict possible social reactions which are reflected in the form of spatiotemporal data.

ACKNOWLEDGEMENTS

The present study is supported by the National Key R&D Program of China (2017YFB0503700 and 2018YFB2100701), the Research Program of Beijing Advanced Innovation Center for Future Urban Design (UDC2019031321), the Pyramid Talent Trai-ning Project of Beijing University of Civil Engineering and Architecture (JDYC20200322), and the National Natural Science Foundation of China (41601389).

REFERENCES

Baker, A. C., Larcker, D. F., Wang, C. C., 2022. How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144(2), 370–395.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1), 249–275.

Chang, M.-C., Kahn, R., Li, Y.-A., Lee, C.-S., Buckee, C. O., Chang, H.-H., 2021. Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. *BMC public health*, 21(1), 1–10.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, AAAI Press, 226–231.

Fang, H., Wang, L., Yang, Y., 2020. Human mobility restrictions and the spread of the novel coronavirus (2019-nCoV) in China. *Journal of Public Economics*, 191, 104272.

Goodchild, M. F., Glennon, J. A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), 231-241.

Horanont, T., Witayangkurn, A., Sekimoto, Y., Shibasaki, R., 2013. Large-scale auto-GPS analysis for discerning behavior change during crisis. *IEEE Intelligent Systems*, 28(4), 26–34.

Huang, J., Wang, H., Fan, M., Zhuo, A., Sun, Y., Li, Y., 2020. Understanding the impact of the COVID-19 pandemic on transportation-related behaviors with human mobility data. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3443–3450.

Huang, W., Li, S., Liu, X., Ban, Y., 2015. Predicting human mobility with activity changes. *International Journal of Geographical Information Science*, 29(9), 1569-1587.

Huang, Z., Chen, Y., Wan, L., Peng, X., 2017. GeoSpark SQL: An effective framework enabling spatial queries on Spark. *ISPRS International Journal of Geo-Information*, 6(9), 285.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y. et al., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Li, Y., Li, M., Rice, M., Zhang, H., Sha, D., Li, M., Su, Y., Yang, C., 2021. The impact of policy measures on human mobility, COVID-19 cases, and mortality in the US: a spatiotemporal perspective. *International Journal of Environmental Research and Public Health*, 18(3), 996.

Mu, X., Yeh, A. G.-O., Zhang, X., 2020. The interplay of spatial spread of COVID-19 and human mobility in the urban system of China during the Chinese New Year. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/2399808320954211.

Terzi, E., Tsaparas, P., 2006. Efficient algorithms for sequence segmentation. *Proceedings of the 2006 SIAM International Conference on Data Mining*, SIAM, 316–327.

Vladimir, E.-C., 2002. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1), 65–75.

Xiong, C., Hu, S., Yang, M., Luo, W., Zhang, L., 2020. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences*, 117(44), 27087–27089.

Yabe, T., Tsubouchi, K., Fujiwara, N., Wada, T., Sekimoto, Y., Ukkusuri, S. V., 2020. Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. *Scientific reports*, 10(1), 1–9.

Yin, F., Lv, J., Zhang, X., Xia, X., Wu, J., 2020. COVID-19 information propagation dynamics in the Chinese Sina-microblog. *Mathematical Biosciences and Engineering*, 17(3), 2676–2692.

Yoo, J. S., Bow, M., 2012. Mining spatial colocation patterns: a different framework. *Data Mining and Knowledge Discovery*, 24(1), 159–194.

Yu, M., Yang, C., Li, Y., 2018. Big data in natural disaster management: a review. *Geosciences*, 8(5), 165.

Zhao, Y., Cheng, S., Yu, X., Xu, H., 2020. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. *Journal of medical Internet research*, 22(5), e18825.

Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., Xia, J., 2020. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health*, 2(8), e417–e424.

APPENDIX

Abbreviations:

BSS	Bike Sharing System
GIS	Geographic Information System
LBS	Location-Based Service
OLS	Ordinary Least Squares
POI	Point Of Interest
VGI	Volunteered Geographical Information
RA	Residential Area
HC	High-tech Company
OC	Ordinary Company
SS	Subway Station
SP	Shopping Plaza
SM	Supermarket
TH	Tertiary Hospital