國立臺灣大學社會科學院經濟學系

碩士論文

Department of Economics

College of Social Sciences

National Taiwan University

Master's Thesis

電信數據之時空分析：居住地遷移與智慧型手機使用對移動行為與通訊模式之影響

Spatio-temporal Analysis of Telecommunications Data: Effects of Residential Shifts and Smartphone Adoption on Mobility Patterns and Communication Behavior

葉秀軒

Hsiu-Hsuan Yeh

指導教授: 謝志昇 博士

Advisor: Chih-Sheng Hsieh, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202502893

Acknowledgements

首先，我要感謝我的父母對於攻讀研究所的理解與支持，並在經濟上給予充分的資助，讓我能夠專心於學業。此外，感謝我的指導教授謝志昇老師，在研究過程中給予寶貴的指導，協助釐清重要的研究主題，熟悉相關文獻，以及耐心傾聽我的想法，並給予我充分的研究自由度。在此也要感謝在台大經濟系六年，所有傳授給我知識的教授們。各位老師不僅為我奠定紮實的學術基礎，也讓我學習到如何品味學術作品以及理解抽象概念，讓我能夠在研究的過程中即使碰到不熟悉的主題，也能夠循序漸進尋找答案。接著，我要感謝我的女朋友以及所有朋友們，在我感到壓力與疲憊時，總是邀請我吃飯和運動，給予我關懷與支持，讓我能夠調適心情，重新投入研究。最後，我也要感謝大型語言模型的出現，特別是 Claude 和 NotebookLM，協助我閱讀文獻、理解數學思路以及修正寫作，大大提升了研究效率與學習品質。再次感謝所有在我求學路上給予幫助的人。

doi:10.6342/NTU202502893

摘要

我們使用大量匿名通話記錄，用以研究居住地變遷與智慧型手機採用對移動與通訊行為的影響。資料涵蓋每月超過十五億通話，涉及超過五十萬個電話號碼，而時間涵蓋 2013 年 8 月至 2014 年 5 月。

我們發現居住地搬遷具有顯著的時間變異效應。遷徙者在搬遷期間傾向於更頻繁地通話，建立更多元的聯絡關係，且主要與原本身處遠距的朋友互動。然而，這些效應會迅速消退回原本水準，或持續發展為負向趨勢，例如互動對象變得較不多元，或聯絡距離縮短。從移動行為的角度來看，搬遷會導致使用者的活動範圍擴大，並出現較難預測的移動模式，儘管這些效應隨時間也會逐漸趨於穩定並變得可預測。

在採用智慧型手機後，移動模式出現明顯（近乎靜態）的上升變化，可能是因為科技在陌生環境中提供協助。例如，移動的不確定性上升，同時出現較明確的方向偏好。

本研究顯示，針對此行為變化，在大規模遷移或手機科技升級的範疇下，政策應更加關注行動與交通建設的需求。

關鍵字：時空分析、電信網路、居住遷移、智慧型手機使用、移動行為、通訊模式

doi:10.6342/NTU202502893

Abstract

We use over 1.5 billion anonymized call records per month spanning from August 2013 to May 2014, where more than 500,000 phone numbers are involved, to study the impacts of residential shifts (events of changing home locations) and smartphone adoption on mobility and communication behaviors.

We find significant time-variant effects for residential relocations. Migrants tend to call more frequently, engage in more diverse contact relationships, and primarily interact with existing distant friends during relocation periods. These effects quickly fade to original levels or continuously evolve toward negative states, such as less diverse interactions or shorter contact distances. From a mobility perspective, residential relocations cause users to have larger exploration areas and highly unpredictable movement patterns, though these effects also shift to more predictable movement over time.

The notable upward shifts (nearly static) in mobility patterns after smartphone adoption are likely due to technological assistance in unfamiliar environments. For example, movement unpredictability increases along with relatively clearer directional preferences.

Our work provides evidence-based policy implications that mobile and transportation infrastructure needs are worth considering during periods when large-scale population displacements or mobile technology upgrades occur.

Keywords: Spatio-temporal Analysis, Telecommunications, Residential Shifts, Smartphone Adoption, Mobility Patterns, Communication Behavior

doi:10.6342/NTU202502893

Contents

doi:10.6342/NTU202502893

doi:10.6342/NTU202502893

doi:10.6342/NTU202502893

List of Figures

doi:10.6342/NTU202502893

doi:10.6342/NTU202502893

List of Tables

doi:10.6342/NTU202502893

Chapter 1 Introduction

We employ call detailed records (CDRs) to study human mobility patterns and mobile communication behavior. Individuals move and interact with others on a daily basis, and CDRs record mobility and communication behaviors at nearly the individual level—that is, observation units are phone numbers rather than actual phone users—facilitating flexible aggregations across multiple hierarchies based on practitioners' needs for answering various research questions. Macro patterns of mobility and communication behaviors revealed in CDRs can provide significant policy implications across multiple domains. For instance, mobility patterns derived from CDRs can inform urban transportation planning by revealing commuting flows (Phithakkitnukoon, Smoreda and Olivier (2012)) and peak travel times (Tongsinoot and Muangsin (2017)), enabling policymakers to optimize public transit routes and schedules. During public health emergencies, CDR mobility data can help design targeted lockdowns by estimating transmission flows (Wesolowski et al. (2016)) or identifying high-risk residential neighborhoods for restrictions while keeping essential services and supply chains operational. Additionally, phone usage patterns can identify digital divides and socioeconomic disparities (Onnela et al. (2007), Blumenstock, Cadamuro and On (2015)) in mobile service usage, helping governments decide where to build better networks and support underserved communities.

We summarize mobility and mobile communication behaviors using six features,

doi:10.6342/NTU202502893

with each behavior characterized by three features. For mobility, we curate radius of gyration (Gonzalez, Hidalgo and Barabasi (2008), Ranjan et al. (2012), Pappalardo et al. (2015)), movement entropy (Eagle, Macy and Claxton (2010), Song et al. (2010), Pappalardo et al. (2016)) and eccentricity (Yuan, Raubal and Liu (2012), Zhao et al. (2019)), which seeks to measure the spatial dispersion of travel, the diversity of human movement, and how closely the spatial distribution of locations resembles an ellipse, respectively. To characterize mobile communication patterns, we construct total call duration, contact entropy (Eagle, Macy and Claxton (2010), Pappalardo et al. (2016)), and contact distance, which quantify the differentials in relationship intensities across contacts, the diversity of mobile interactions, and the spatial reach of social interactions, respectively.

While we did not invent these features, we improved them by incorporating the load sharing mechanism that is prevalent in the telecommunication industry, containing two components: load balancing (Ayesha et al. (2019)) and handover (Márquez-Barja et al. (2011)). Two issues arise with the load sharing mechanism. First, the telecom base station handling a call event may not always be the closest one to the user (Yuan, Raubal and Liu (2012)). Second, a single call communication can generate multiple call records due to the change of telecom base station processing that call. Ignoring the load sharing mechanism can introduce biases when inferring users' significant locations based on the staying frequency and the intensities of social ties characterized by mobile communication. Therefore, we make methodological contributions by proposing that the randomness of mobile interactions should be modeled using relative call duration rather than the number of calls, and that the randomness of location stays should be modeled using the relative number of days a particular base station handles calls rather than the total number of call events it processes. Besides, we also apply DBSCAN, a machine learning clustering algorithm, to

doi:10.6342/NTU202502893

mitigate localization errors (Ayesha et al. (2019)), thereby improving the robustness of home location estimation.

CDRs are already heavily utilized in human mobility research (Gonzalez, Hidalgo and Barabasi (2008), Song et al. (2010), Wesolowski et al. (2016)) and social network analysis (Onnela et al. (2007), Cho, Myers and Leskovec (2011), Barwick et al. (2023)). However, the majority of research focuses on modeling statistical properties of these behavioral features or examining correlations between social networks and mobility. We delve into a novel research topic that examines micro-level interactions instead of inspecting correlations between macro patterns while still providing aggregate implications. Specifically, we identify significant treatments that substantially influence mobility and communication behaviors, followed by treatment effect identification that examines how effects on behavioral features unfold over time through a difference-in-differences (DiD) design with multiple periods and variation in treatment timing.

The two treatments are residential shifts and smartphone adoption, and the treatment effect dynamics are estimated through an approach proposed by Callaway and Sant' Anna (2021), which is robust to heterogeneous effects over time and across treatment-timing groups. We found that there is a temporary surge in total call duration and contact distance during the month of relocation, arising from migrants' attempts to contact geographically distant social connections. Moreover, during this same period, migrants engage in more diversified social interactions, however, following the completion of relocation, migrants tend to spend less time on mobile communication with less diversified interactions. On the other hand, radius of gyration substantially increases contemporaneously with the completion of residential relocation, while the effect quickly fades in the following months. Mobility characteristics transition from highly unpredictable spatial appearances with spa-

doi:10.6342/NTU202502893

tial stretching along a fixed direction to predictable patterns with roughly circular spatial distribution. The effects of smartphone adoption are nearly constant over time, and the changes are positive. The most notable influence is the increase in total call duration, while movement entropy, eccentricity, and contact entropy all show modest increases.

The estimation results help interpret anomalous events when using CDRs to monitor mobility and communication behaviors during periods or in regions experiencing large immigrant influxes or significant technology adoption, both commonly seen in developing countries. Moreover, our work also suggests that policy awareness should increase regarding the need for mobile and transportation infrastructure when significant population displacement or mobile technology updates occur. Examples of population displacement include refugee resettlement programs (e.g., around 1 million Syrian refugees who fled civil war and resettled in Germany during 2015-2016), natural disaster relocations (e.g., about 15 million people resettled within China following the 2008 Wenchuan earthquake), or environmental displacement due to industrial pollution (e.g., 833 families relocated from Love Canal, New York during 1978-1980 due to toxic chemical contamination). Mobile technology updates contain network infrastructure upgrades (2G to 3G to 4G to 5G), and GPS-enabled services adoption.

The remaining content is structured as follows. Chapter two provides a literature review on internal migration and estimating home locations and identifying residential shifts through CDRs. Chapter three introduces our data sources, how we identify residential shifts and smartphone adoption, and how various behavioral features are constructed. Chapter four explains how we assure the existence of anticipation and numerous intuitions from the estimation results. Finally, Chapter five delves into the summary, limitations, and future work of this study.

doi:10.6342/NTU202502893

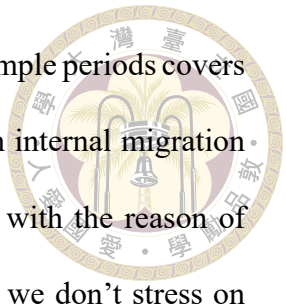
Chapter 2 Literature Review

2.1 Residential Shifts and Internal Migration

CDRs is a collection of geotagged phone call records (Table 3.1) over a period of time. Although the observation units are phone numbers rather than actual phone users, we will proceed to use phone users as the sample units since, on average, people have 1.2 phone devices. By looking at individual level, we can trace phone users' geographical appearances with temporal dimension. Based on the spatial patterns distributed on the map, we can estimate their significant locations, such as their residential coordinates and workplace location. Besides, the temporal dimension offers a particularly exciting opportunity: we can identify events of residential shifts, depicting migration flows that were previously impossible to capture at such scale and precision through traditional census-based survey.

This type of human movement is referred to as internal migration, unlike international migration, and it signifies population flow that occurs within a country. Internal migration is a phenomenon that has captured economists' attention for over a century. This strand of literature often involves modeling migration decisions (Hunt and Mueller (2004), Espíndola, Silveira and Penna (2006), Wang-Lu and Valerio Mendoza (2023)) or inspecting the impacts of migration on the destination region (Boustan, Fishback and Kantor (2010), Bryan and Morten (2019), Imbert et al. (2022)).

doi:10.6342/NTU202502893



Our CDRs are collected from Sichuan province, China, and the sample periods covers from August 2013 to May 2014. In China, most of the discussion on internal migration is concentrated on the rural-urban migration on the prefecture level with the reason of finding better jobs or seeking education opportunities. Nonetheless, we don't stress on any particular context of internal migration and focus on the detection of residential shifts and their influences on human behaviors.

One of our methodological contributions is identifying large-scale inter-prefecture migration flows through novel data sources. Specifically, we employ CDRs and develop a systematic pipeline with two stages to identify individuals changing their residential locations, revealing the internal migration flows across prefectures. Utilizing CDRs to identify internal migration flow has several advantages. First, CDRs capture mobility patterns for virtually all mobile phone users in a region, including populations often underrepresented in conventional surveys such as transient residents, undocumented individuals, and those reluctant to participate in formal governmental data collection. Second, CDRs provide continuous temporal coverage rather than the periodic snapshots offered by censuses, enabling the detection of short-term or seasonal relocations. Third, this approach is also cost-effective compared to the large amount of money and human resources devoted to completing a census, as telecommunication companies automatically collect these phone records for billing purposes.

2.2 Home Location Estimation through CDRs

It is important to clarify that the coordinates attached to each phone call record don't precisely represent the exact geographical position of either the caller or callee. Rather,

doi:10.6342/NTU202502893

they are the coordinates of the telecom base station handling the call, which serve as a spatial approximation for the phone user's position while initiating or receiving phone calls. The approximation is not always accurate due to the load balancing mechanism.

Most of the studies estimate home locations through CDRs by selecting the locations of the telecom base station that handles the call events most frequently over the whole sample period (Cho, Myers and Leskovec (2011), Phithakkitnukoon, Smoreda and Olivier (2012)), weekly (Barwick et al. (2023)), or monthly (Phithakkitnukoon (2022)) from the nighttime call records. This simple approach seems to be acceptable for people who have a large amount of phone call records. However, for those who have limited observations of call events, the simple approach is not reliable. A more robust approach would be running a spatial clustering algorithm over a set of telecom base stations' locations (Isaacman et al. (2011), Yang et al. (2014), Ayesha et al. (2019)). As mentioned, our estimation strategy of residential location encompasses two stages, and the first stage recognizes the clustered patterns and leverage DBSCAN (Ester et al. (1996)), a renowned machine learning algorithm in the clustering domain, to uncover them. Note that Ayesha et al. (2019) also leveraged DBSCAN to estimate home locations.

DBSCAN's flexibility has made it a popular choice for analyzing spatial patterns (Yang et al. (2014), Shi et al. (2014), Domínguez et al. (2017)) and mobile communication behaviors (Karahoca and Kara (2006), Jabbar and Suharjito (2020)). We inherit this idea, including it in our two-staged approach, which carries the specific goal of identifying residential shifts. Our approach makes this identification feasible by incorporating temporal information following the spatial clustering process. Even more recent developments in significant location inference (Tongsinoot and Muangsin (2017), Luo et al. (2020)) don't explicitly consider the situation where people might change their home locations.

doi:10.6342/NTU202502893

2.3 Detection of Residential Shifts through CDRs

Several studies have already taken advantage of CDRs to identify residential relocation. Although most of the works on home location inference are not designed for detecting home location shifts in that they often estimate one location over the whole sample period. However, we can still apply these approaches on multiple fixed-size time window (Blumenstock (2012), Phithakkitnukoon (2022), Blumenstock, Chi and Tan (2025)), e.g., daily, weekly, or monthly, or between two time periods (Lai et al. (2019), Dias et al. (2022)) and if more than one home location is found, we can consider it as a residential shift. Phithakkitnukoon (2022) is the case where they apply the simple approach on each month to detect residential shifts while Dias et al. (2022) adopt Isaacman et al. (2011)'s home location estimation method on January to March 2013 and July to September 2013, respectively. This strategy for inferring residential shift requires several predefined parameters, such as the minimum time span of each residential location to exclude short-term visits or distance threshold to define the separation of two home locations. Hence, a heavy procedure of sensitive analysis is required to select the appropriate parameters.

Büchel et al. (2020) also utilize CDRs to identify residential shift and benefits from high-quality data billing addresses, making residential estimation extremely precise and identification of migrants is simply based on whether individuals change their residential locations. However, as such data is not available in most scenarios, thereby failing to be widely applicable.

Chi et al. (2020) is a closely related work. They abandon the first stage by clustering coordinates of telecom base stations if they are in the same administrative district (e.g., prefectures in China). Furthermore, they apply the clustering algorithm on time axis for

doi:10.6342/NTU202502893

each district to identify contiguous time segments and further merge segments tailored to a particular district if no other segments from the other district are found in the same time window defined by the target merging segments. At the last step, they allow for overlap between segments from different districts to form home locations at different periods.

Our two-stage approach provides fine spatial resolution through first-stage spatial clustering, though this isn't strictly necessary for cross-district migration flows. The second stage uses temporal filtering—simpler than temporal clustering—that identifies home clusters by assuming the largest cluster is most likely to be one of the "home" clusters as people won't change their home locations too often. However, their second stage may be more robust as it allows for overlap between clusters.

Our approach stands out for being universal, efficient, and comprehensive. It relies solely on CDRs, requiring no additional information. There is only one parameter: the maximum distance between two locations to be considered neighbors, and since we follow the literature on internal migration in China by focusing on the inter-prefecture migration flows, sensitivity analysis of this parameter is not strictly necessary. Furthermore, we unveil inherent spatial-temporal patterns by utilizing the unsupervised clustering algorithm and maximally exploiting the full range of temporal information.

doi:10.6342/NTU202502893

Chapter 3 Data and Methods

3.1 Datasets

Table 3.1: An example of geotagged CDRs

Client Number	Duration	Start Time	End Time	Calling Number	Called Number	Cell ID
66ak2s5v	62.0	2013-08-31 09:32:08	2013-08-31 09:33:10	66ak2s5v	moyl2k57	3649
ltzkksuv	148.0	2013-08-31 09:33:55	2013-08-31 09:36:23	ltzkksuv	hjo0ksut	3B56
njo45k8v	46.0	2013-08-31 09:36:03	2013-08-31 09:36:49	8yro82d5	njo45k8v	394C
			⋮			

Notes: All phone numbers are anonymized. Cell IDs are the IDs of the telecom base stations handling call events for client numbers, which are either calling numbers or called numbers. We have a dataset which records the geographical coordinates of cell IDs.

We have three datasets: CDRs, coordinates of telecom base stations, and one that contains phone users' profile features, such as age, gender, phone brand, service type, etc. Three datasets are provided by China Telecom, one of the three telecommunications service providers in China. Call records are located in Sichuan Province, and the majority of them are located in Deyang Prefecture. By joining call records with coordinate data on cell IDs and then grouping by user phone numbers, we can obtain a set of geographic locations associated to each phone user from August 2013 to May 2014, which serve as the location approximation of users. User profile data is uncommonly available due to privacy issues but critical for selecting samples of interest in CDR-based analysis. When constructing various mobility and mobile communication features or estimating residen-

doi:10.6342/NTU202502893

tial coordinates, we want the samples to be consistently observable throughout the whole sample period, i.e., they don't cancel or register service in the middle of the sample period, to avoid spurious analysis due to too many missing records.

The cleaning of CDRs involves removing records where (i) both the calling or called number is not mobile phone numbers, (ii) both the start or end time is not valid timestamps, and (iii) the duration of calls is less than or equal to 0. For user profile data, we remove client numbers in profile data whose ID card numbers are not in the correct specification. After the cleaning, we have 0.5 to 0.6 billion call records per month and 348,241 client numbers whose profile features are available in each month.

3.2 Notations

Denote V as a set of the phone users, B as a set of telecom base stations and $T = M \times D \times H$ as a set of timestamps where M is a set of calendar months spanning from Aug 2013 to May 2014, $D := \{1, 2, ..., 31\}$ and H is a set of all possibles times in a day.

Definition 3.2.1 (Call Detailed Records)**.** CDRs denoted by R is a collection of phone calls, which are 4-tuples, containing information of the caller, recipient, timestamp, and telecom base station that services the call. It's defined as:

$$R := \{(i, j, t, b) \in V \times V \times T \times B \mid i \text{ calls } j \text{ at timestamp } t \text{ and the call is serviced by } b\}.$$

Definition 3.2.2 (A User's Nighttime Call Records Serviced by a Telecom Base Station)**.** Given CDRs R, we can filter call events that are either made or received by a user i and serviced by a telecom base station b during nighttime. Note that the definition of nighttime

doi:10.6342/NTU202502893

follows Barwick et al. (2023). It's denoted by $R_{i,b}^{\text{night}}$ and defined as:

$$R_{i,b}^{\text{night}} := \{r \in R \mid \text{there exists } t \in T \text{ where } t \geq 10 \text{ p.m. and } t \leq 7 \text{ a.m.} \\ \text{and } j \in V \text{ such that } r \in \{(i,j,b,t),(j,i,b,t)\}\}.$$

Definition 3.2.3 (A Set of Timestamps of a User's Nighttime Call Records Serviced by a Telecom Base Station)**.** A subset $T_{i,b}^{\text{night}} \subset T$ is a collection of timestamps associated to $R_{i,b}^{\text{night}}$, and it's defined as:

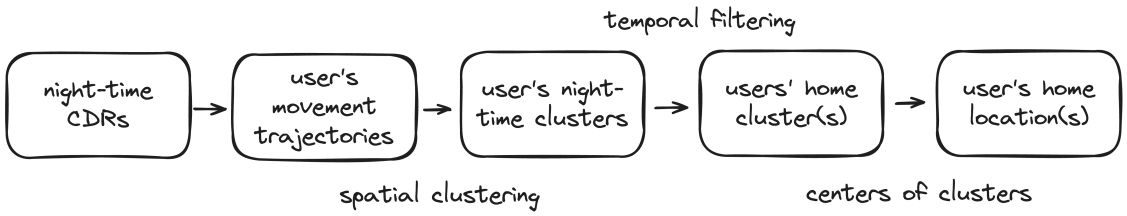
$$T_{i,b}^{\text{night}} := \{t \in T \mid \text{there exists } j \in V \text{ such that either } (i,j,t,b) \text{ or } (j,i,t,b) \in R_{i,b}^{\text{night}}\}.$$

Definition 3.2.4 (A Subset of Telecom Base Stations Connected to a User during Nighttime)**.** A subset $B_i^{\text{night}} \subset B$ of telecom base stations connected to a user $i \in V$ during nighttime is defined as:

$$B_i^{\text{night}} := \{b \in B \mid \text{there exists } j \in V \text{ and } t \in T \\ \text{such that either } (i,j,t,b) \text{ or } (j,i,t,b) \in R_{i,b}^{\text{night}}\}.$$

3.3 Home Location Estimation

Figure 3.1: Pipeline for Estimating A Phone User's Home Locations



This section aims at introducing how phone users' home locations are estimated, and Figure 3.1 presents all the steps for completing the task. Before directly diving into how

doi:10.6342/NTU202502893

our proposed estimation works, we provide a brief preliminary introduction to the issues and limitations of using CDRs to locate user locations, which motivates the use of spatial clustering.

To establish our terminology, we use ”call event” to refer to mobile communication established through a single phone call. However, a call event can generate multiple call records due to changes in telecom base stations resulting from load sharing mechanisms. We term this mismatch between call events and call records the ”load sharing effect” (Ayesha et al. (2019)). Throughout this text, we use the terms ”telecom base station”, ”telecom station”, and ”base station” interchangeably, all referring to the mobile infrastructure that routes phone calls.

The load sharing mechanism is triggered under two primary circumstances: when there is a need to balance traffic among adjacent base stations (load balancing) or when a phone user moves across different service areas of base stations (handover). We redefine the ”load sharing” concept introduced in Ayesha et al. (2019), as their definition appears more similar to ”load balancing”. We propose that ”load sharing” should encompass a broader scope, including the sharing mechanisms observed in handover scenarios. Since our analysis focuses on non-sequential mobility patterns—prioritizing spatial characteristics of where users stay or visit rather than temporal movement sequences—the load balancing aspect becomes particularly relevant.

A direct way to characterize the load balancing is that the telecom station handling a call record is not necessarily the closest one to the actual location where the call event occurs. It is important to note that ”locations” refer to geographic coordinates throughout this text. Specifically, stay locations refer to the exact location where a call event

doi:10.6342/NTU202502893

is initiated. This location is unobservable and differs from all locations of base stations associated with the call records relevant to that call event. We refer to the base station locations as ”observed locations,” which serve as spatial approximations of the true stay locations. Multiple observed locations can be associated with a single stay location through two mechanisms: (i) load balancing (as discussed above) and (ii) repeated usage patterns, where users frequently make calls from the same stay location, generating multiple observed locations that correspond to the same underlying stay location.

From an analytical standpoint, characterizing the mobility behavior revealed in CDRs requires reconstructing stay locations from observed locations. This reconstruction challenge, combined with our focus on spatial characteristics rather than sequential movement patterns, naturally motivates the application of spatial clustering techniques to group nearby observed locations.

Home location estimation represents a specific application of stay location reconstruction, where we focus on identifying residential places using exclusively nighttime call records. During nighttime hours, users may have multiple stay locations, and spatial clustering allows us to identify sets of observed locations that form nighttime clusters around potential home location candidates. The challenge then becomes filtering true ” home clusters” from the broader set of nighttime clusters, which we accomplish through a temporal filtering scheme introduced in Section 3.3.2. Our proposed approach therefore operates in two stages: spatial pattern aggregation followed by temporal filtering.

doi:10.6342/NTU202502893

3.3.1 Spatial Clustering

Our choice of spatial clustering algorithm is DBSCAN and we apply DBSCAN to each phone user over a set of observed locations.[1] As mentioned, observed locations refer to base station coordinates and connect to an injective function loc : $B \to \mathbb{R}^2$ which maps each base station $b \in B$ to its geographic coordinates $\text{loc}(b) = (lon_b, lat_b)$ where lon_b and lat_b represent the longitude and latitude of telecom base station b, respectively. Our choice of DBSCAN is based on the fact that it does not require predefining the number of clusters, unlike K-means, offering greater flexibility in identifying natural clusters of observed locations that correspond to underlying stay locations.

Recall that multiple observed locations can be associated with a single stay location through two mechanisms: (i) load balancing effects where the handling base station is not the closest to the call records and (ii) repeated usage patterns where users frequently make calls from the same stay location. Spatial clustering leverages this relationship by grouping nearby observed locations that likely correspond to the same underlying stay location. For nighttime clusters specifically, each cluster represents a set of observed locations that corresponds to a single home location candidate (the unobservable true stay location).

Definition 3.3.1 (A Nighttime Cluster for a User)**.** The k-th nighttime cluster $C_{i,k}^{\text{night}}$ is obtained through the application of DBSCAN, and defined by:

$$C_{i,k}^{\text{night}} := \{b \in B_i^{\text{night}} \mid \text{for all } b_m, b_n \in C_{i,k}^{\text{night}} \text{ where } m \neq n, \\ (\text{loc}(b_m), \text{loc}(b_n)) \text{ satisfy DBSCAN clustering}\}$$

[1] We apply scikit-learn's implementation of DBSCAN to each phone user.

doi:10.6342/NTU202502893

Note that, we set the two parameters of DBSCAN: *min_samples* and *eps* to 1 and 5 (km), respectively. For the discussions on the parameter choices, please refer to the section of Parameter Choices of DBSCAN. Besides, we further denote $C_i^{\text{night}} = \{C_{i,1}^{\text{night}}, C_{i,2}^{\text{night}}, \cdots\}$ as a set of nighttime clusters for user i.

3.3.2 Temporal Filtering

After the procedure of spatial clustering, we can zoom out the level of spatial analysis from locations to clusters so the issues arising from the load sharing mechanism are now resolved. Therefore, we define the new notations for both subsets of CDRs and timestamps based on C_i^{night}, which previously defined while considering B_i^{night}.

Definition 3.3.2 (A User's Nighttime Call Records Serviced by a Nighttime Cluster)**.** For a cluster $C_{i,k}^{\text{night}} \in C_i^{\text{night}}$, a set of call records serviced by $C_{i,k}^{\text{night}}$ is defined as:

$$\mathcal{R}_{i,k}^{\text{night}} := \bigcup_{b \in C_{i,k}^{\text{night}}} R_{i,b}^{\text{night}}.$$

Definition 3.3.3 (A Set of Timestamps of a User's Nighttime Call Records Serviced by A Nighttime Cluster)**.** For a cluster $C_{i,k}^{\text{night}} \in C_i^{\text{night}}$, the timestamps of $\mathcal{R}_{i,k}^{\text{night}}$ is defined as:

$$\mathcal{T}_{i,k}^{\text{night}} := \bigcup_{b \in C_{i,k}^{\text{night}}} T_{i,b}^{\text{night}}.$$

The second stage of the home location estimation is to apply the temporal filtering trick, aiming at obtaining home clusters from a set of nighttime clusters, and the home clusters are leveraged to estimate users' home locations. Furthermore, the purpose is to keep those clusters that are temporal representative, processing substantial calls, which are evaluated through the measure of "temporal size".

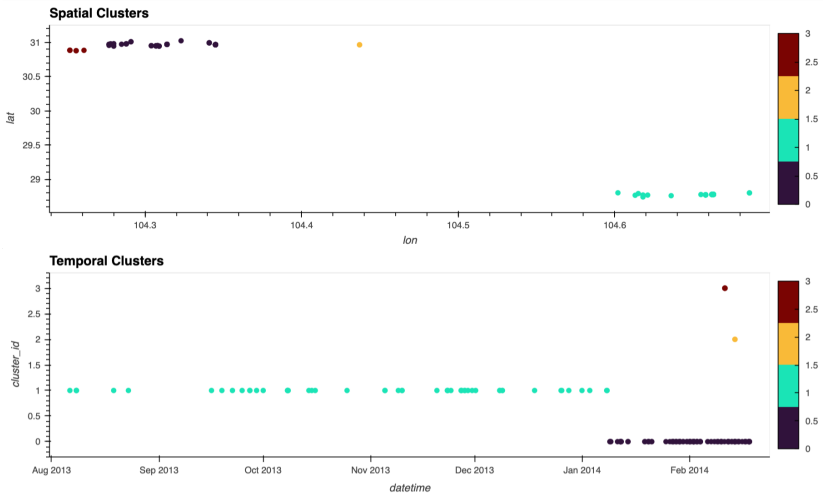
doi:10.6342/NTU202502893

Function 3.3.1 (A Measure of Temporal Size of a Set of Timestamps)**.** The measure of temporal size of a subset $\tilde{T} \subset T$ of timestamps is defined as a measure $\mu : 2^T \to \mathbb{N}$ where

$$\mu(\tilde{T}) = |\{(m, d) \in M \times D \mid \text{there exists } h \in H \text{ such that } (m, d, h) \in \tilde{T}\}|. \quad (3.1)$$

Let us elaborate more on the definition of a representative cluster. We define a cluster to be representative if the temporal spans associated with each cluster are non-overlapping. An example can be found in Figure 3.2 where the 0-th cluster (black) doesn't temporally overlap with the 1-st cluster (green) in the lower plot. Therefore, each home cluster should correspond to a separate life period where the person lived in a particular location. Moreover, the identified home clusters are temporally sequential and mutually exclusive–when one residential period ends, the next begins, with no overlap between them. This somehow defines the temporal clusters and creates a clear timeline of residential history where each home cluster represents a distinct "home era" in chronological order.

Figure 3.2: Visualization of Proposed Two-stage Home Locations Estimation



Notes: The upper plot's x and y axis represent the longitude and latitude, respectively. The lower plot shows the temporal span of each cluster where the x axis represents the timestamp and the y axis represents the cluster index. The point is colored by the cluster index. In this plot, the user are identifed to change the home location within green cluster (cluster index 1) to the black one (cluster index 0).

doi:10.6342/NTU202502893

Before defining the overlap, we need to define objects associated to two distinct clusters that can be identified to be overlapping with each other. The object is the service time intervals, which are continuous time intervals starting from the first time that a cluster services calls and ending at the last time that a cluster services calls.

Function 3.3.2 (The First & Last Timestamp of a Set of Timestamps)**.** Consider a set of timestamps $\tilde{T} \subset T$, the first and the last timestamps of which can be obtained by the following functions.

$$t^{\text{first}} : 2^T \to T, \quad t^{\text{first}}(\tilde{T}) = \min_t \{t \in \tilde{T}\}.$$

$$t^{\text{last}} : 2^T \to T, \quad t^{\text{last}}(\tilde{T}) = \max_t \{t \in \tilde{T}\}.$$

Definition 3.3.4 (A Nighttime Cluster's Service Time Interval)**.** A continuous time interval associated to a nighttime cluster $C_{i,k}^{\text{night}} \in C_i^{\text{night}}$ is defined as: $[t^{\text{first}}(\mathcal{T}_{i,k}^{\text{night}}), t^{\text{last}}(\mathcal{T}_{i,k}^{\text{night}})]$, which represents that the cluster $C_{i,k}^{\text{night}}$ services nighttime phone calls for client i during this period.

Definition 3.3.5 (Temporal Overlappness between Two Nighttime Clusters)**.** Consider two nighttime clusters, $C_{i,1}^{\text{night}}, C_{i,2}^{\text{night}}$, they overlap with each other temporally means there exists $t \in T$ such that t is in both service time intervals of $C_{i,1}^{\text{night}}$ and $C_{i,2}^{\text{night}}$.

With the definition of overlap, our temporal filtering trick is a sequential process, where we first sort the clusters in C_i^{night} in descending order by the temporal sizes defined by $\mu(\mathcal{T}_{i,k}^{\text{night}})$ where $C_{i,k}^{\text{night}} \in C_i^{\text{night}}$, and then we iteratively select the cluster to be the home cluster that is not temporally overlap with any of the pre-selected home clusters. The algorithm is shown in Algorithm 1. Note that at the last step of the algorithm, the home clusters' start times of service time intervals are well-defined based on the corresponding

doi:10.6342/NTU202502893

nighttime clusters.

Given the home clusters, we can define the home locations as the centroids of clusters, computed by the weighted average of the locations of the base stations in the cluster, where the weights are the temporal sizes. Isaacman et al. (2011) also adopt this weight definition while they applied the clustering algorithm designed by Hartigan (1975). They argue that using the number of days instead of the number of calls can reduce the influence of base stations that were only used for a few days but had a burst of activity on those days. For example, during a temporary vacation, people might make many phone calls to share their experiences with their contacts.

Definition 3.3.6 (A User's Home Location)**.** For a user $i \in V$, the l-th home location $(r_c)_{i,l}^{\text{home}}$ is defined by the weighted average over locations $\{\text{loc}(b)\}_{b \in C_{i,l}^{\text{home}}}$ of telecom base stations that are contained in the l-th home cluster $C_{i,l}^{\text{home}}$ where the weights are defined based on the temporal sizes.

$$(r_c)_{i,l}^{\text{home}} = \sum_{b \in C_{i,l}^{\text{home}}} \frac{\mu(T_{i,b}^{\text{night}})}{\sum_{b \in C_{i,l}^{\text{home}}} \mu(T_{i,b}^{\text{night}})} \text{loc}(b)$$

doi:10.6342/NTU202502893

3.4 Identification of Residential Shift and Its Timing

Residential shift is identified if a user's home location changes. Since CDRs record users' locations discretely and irregularly, it's ambiguous to decide the migration timing, as there may be latency between the actual date where users relocate and the date of the first timestamp where the second home cluster serves the calls. The determination on the migration timing is crucial in the design of the DiD with multiple periods as it determines the pre-treatment and post-treatment periods where the parallel trend assumption is tested, and the treatment effect dynamics are inspected, respectively.

At first thought, we can define the migration timing as (i) the date of the last timestamp where the first home cluster serves the calls, (ii) the date of the first timestamp where the second home cluster serves the calls, or (iii) the date in the middle of them. We opt for a conservative approach and select the second option, which guarantees that the chosen date occurs after migration has already been completed. This decision results in the necessity to consider the violation of "no anticipation" (Callaway and Sant' Anna (2021), Sun and Abraham (2021), Borusyak, Jaravel and Spiess (2024)) because people might start to collect information for better preparation before migrating to another prefecture, causing the divergent paths of mobility and mobile communication features between migrants and non-migrants before a actual migration event takes place. Note that the no anticipation assumption doesn't require hold in all pre-treatment periods; instead, it's plausible to assume it holds until a period before the treatment.

We are discussing treatment effect dynamics on a monthly level from August 2013 to May 2014. Therefore, to establish clean notations, we index the monthly periods from integer 1 to 10 where August 2013 corresponds to 1, September 2013 corresponds to 2,

doi:10.6342/NTU202502893

and so on. Furthermore, we denote $\mathcal{M} := \{1, \cdots, 10\}$ as the set of monthly periods after indexing. Below we define the treatment group consisting of four subgroups $\{\mathcal{G}_g\}_{g=4,5,6,7}$ as migrants who relocate in different months g and define the never-treated group $\mathcal{G}_\infty$ as non-migrants. We utilize the symbol ∞ to indicate a non-migrant will relocate in the far-away future, i.e., infinity period (Sun and Abraham (2021), Borusyak, Jaravel and Spiess (2024)).

Definition 3.4.1 (Migrant)**.** A phone user $i \in V$ is a migrant associated with the group $\mathcal{G}_g$ if (i) user i only changes home locations once during the sample period, (ii) the two home locations are in different prefectures, and (iii) the migration event occurs in month $g \in \{4, 5, 6, 7\} \subset \mathcal{M}$.

Definition 3.4.2 (Non-Migrant)**.** A phone user $i \in V$ is a non-migrant associated with the group $\mathcal{G}_\infty$ if i doesn't change the home locations throughout the sample period.

Applying these definitions, we identify 1,274 migrants and 291,465 users who do not shift their residential locations throughout the sample period. For the detailed explanations on how criterions are set, please refer to the Section E.3 We require migrants to have only relocated once because among users who have multiple residential shifts, most of them have only gone through it once, accounting for 99.65%. Besides, longer distance of residential movements should be more likely to have substantial impacts on mobility and mobile communication patterns so we restrict our discussions on inter-prefecture migrants. As Table E.3 demonstrates, we don't lose too many migrants by restricting the definition of migrants to those crossing prefectures.

doi:10.6342/NTU202502893

3.5 Detection of Smartphone Adoption

Aside from residential shifts, our user information data enables us to inspect the effects of other shift events, specifically upgrading to smartphones. Smartphones have abundant functionalities that facilitate mobile communication and better assist users in exploring unfamiliar environments through GPS technology, potentially altering smartphone adopters' mobility patterns. To study the smartphone-adoption shift, we define the treated and untreated units as follows.

Definition 3.5.1 (Smartphone Adopter)**.** A phone user $i \in V$ is a smartphone adopter associated with group $\mathcal{G}_g$ if user i (i) changes from a non-smartphone device to a smartphone model, (ii) is not observed to switch back to non-smartphone devices throughout the sample periods, and (iii) the adoption event occurs in month $g \in \{4, 5, 6, 7\} \subset \mathcal{M}$, where these months represent the middle of the sample period.

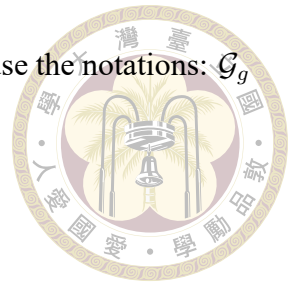
This definition rules out phone users who have multiple cellphones, and constantly switch between smartphone and non-smartphone devices. As defining migrants, we require the events to happen in the middle of the sample period to confirm the reliability of the adoption event, avoiding the observation window bias.

Definition 3.5.2 (Non-Smartphone User)**.** A phone user $i \in V$ is a non-smartphone user associated with group $\mathcal{G}_\infty$ if user i consistently uses non-smartphone devices throughout the sample periods.

Applying these definitions, we identify 81,949 users who consistently use non-smartphone devices throughout the sample period and 9,497 users who change from non-smartphone to smartphone devices. Additionally, 17,717 users are identified as having used both de-

doi:10.6342/NTU202502893

vice types, while 203,776 users consistently use smartphones. We reuse the notations: $\mathcal{G}_g$ and $\mathcal{G}_\infty$ to denote the treated and non-treated cohorts.



3.6 Construction of Outcome Variables

Our outcomes of interest contain two groups: mobile communication and human mobility features. For each group, we craft three kinds of features to study the effects of residential shift and smartphone adoption on these outcome variables.

Mobile communication network features are derived from the reciprocated network and include three kinds of features: call duration, contact entropy, and contact distance. The contact distance isn't commonly seen in the related literature, but it contains rich interpretations of how phone users' mobile communication network geographies look. Besides, it's more complicated to construct compared to the other two as it can't be directly derived from the call detailed records. Instead, we need a preliminary home-estimation procedure, then compute the average geographical distance between a phone user's home and the other contacts' homes. Therefore, it serves as a great extension of our robust home estimation method presented in Section 3.3.

Mobility features consist of radius of gyration, movement entropy, and eccentricity. Jointly considering movement entropy and eccentricity offers additional insights, beyond the "unpredictability" provided by movement entropy alone. Specifically, we can identify distinct mobility patterns, such as whether users exhibit highly random movements that are stretched along one direction (high entropy, high eccentricity) versus more predictable movements that spread evenly across all directions (low entropy and low eccentricity). We will see interesting evolution in these patterns after migration and upgrading to smartphone

doi:10.6342/NTU202502893

devices.

What's more, entropy-based variables can be viewed as measures of diversity, in addition to unpredictability. Employing entropy as a measure of diversity captures another aspect beyond quantity increases, e.g., interacting with more friends or visiting more places. Entropy says diversity should also be credited to the randomness. That is, calling the same number of friends with equal frequency demonstrates a sign of diversity. Besides, we can measure the diversity while eliminating the variation in quantity by normalizing it (dividing by $\log(N)$, where N can be the total number of contacts/locations).

3.6.1 Mobile Communication Network Features

We haven't formally defined what the mobile communication network is. Basically, the nodes of the network are defined as phone users and edges are defined as reciprocated calls that occur on weekdays, which are further defined as follows.

Definition 3.6.1 (A Reciporated Call)**.** A call record $r \in R$ where $r = (i, j, t, b)$ for some $i, j \in V, t \in T$ and $b \in B$ is reciprocated if there exists $t' \in T$ and $b' \in B$ such that $(j, i, t', b') \in R$.

Note that the mobile communication network is a kind of directed network—i.e., for example, a user i calls j and j calls i are considered as two edges, whereas this would be considered as a single edge in an undirected graph. Besides, the edges are weighted by the underlying call duration (in minutes). Since mobile communication is directed, we can naturally build relevant features based on distinct communication directions. To avoid redundantly defining the same features for both incoming and outgoing calls, we will illustrate the feature definitions using outgoing calls as examples.

doi:10.6342/NTU202502893

After constructing the mobile communication network by month, directional call duration for a user $i \in V$ in a given month $m \in \mathcal{M}$ is computed by separately aggregating incoming and outgoing calls' duration. Contact entropy is constructed based on the Shannon entropy normalized by the logarithm of the number of contacts for a user i in a given month m. Normalization is necessary as it accounts for differences in network size across phone users, thereby leading to clearer interpretation of the regression coefficients. Intuitively, increasing unnormalized entropy by 0.3 doesn't mean the same thing for a user with 10 friends versus one with 50 friends.

Definition 3.6.2 (Outgoing Contact Entropy)**.** The outgoing contact entropy $ce^{\text{out}}_{i,m}$ (where ce stands for contact entropy) for user i in month m is defined by:

$$ce^{\text{out}}_{i,m} = \frac{-\sum_{j \in V^{\text{out}}_{i,m}} \hat{p}_{i,j,m} \log(\hat{p}_{i,j,m})}{\log(|V^{\text{out}}_{i,m}|)}$$

where

$$V^{\text{out}}_{i,m} := \{j \in V \mid \text{user } i \text{ has once called user } j \text{ in month } m\} \tag{3.2}$$

and

$$\hat{p}_{i,j,m} = \frac{w_{i,j,m}}{\sum_{k \in V^{\text{out}}_{i,m}} w_{i,k,m}} \tag{3.3}$$

with $w_{i,j,m}$ being the weight of edge from user i to user j in month m. Note that the weight is defined as the total call duration rather than number of phone calls.

Employing call duration as weights for social ties is not a common practice, and Eagle, Macy and Claxton (2010) and Pappalardo et al. (2016) defined contact entropy based on the number of phone calls. Nevertheless, as previously mentioned, we recognize that due to load sharing mechanisms, phone calls might change their served base station, thereby generating multiple call records when actually only a single call is taking place.

doi:10.6342/NTU202502893

Figure 3.3: Pipeline of Constructing Contact Distance

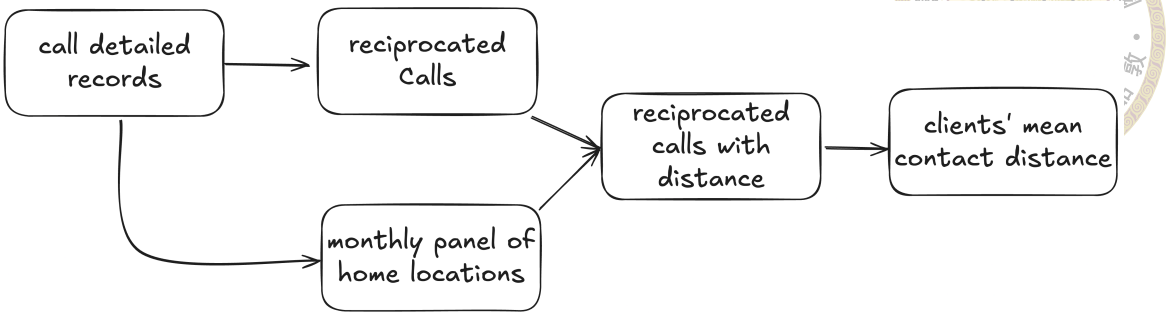


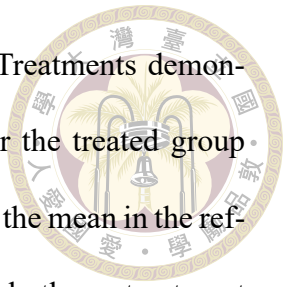
Figure 3.3 demonstrates how to construct the contact distance for user i in month m. Note that the estimation of home location is necessary in our situation as we do not have the geographical locations of both the caller and recipient for each call; instead, only a single location is attached to each phone call—specifically from one telecom base station (see Table 3.1). Therefore, we impute these locations using the corresponding home locations obtained through our proposed estimation approach. Furthermore, as the home location for a given month is estimated through CDRs, the estimation will fail for some months that lack CDRs. We employ forward imputation followed by backward imputation to address this issue. Forward imputation means using month $m-1$ to impute month m's home location (if month m doesn't have any CDR), while backward imputation uses month $m+1$ to impute month m's home location.

Definition 3.6.3 (Outgoing Contact Distance)**.** The outgoing contact distance $cd^{\text{out}}_{i,m}$ (where cd stands for contact distance) for user i in month m is defined by:

$$cd^{\text{out}}_{i,m} = \sum_{j \in V^{\text{out}}_{i,m}} \hat{p}_{i,j,m} \cdot d(\text{home}_{i,m}, \text{home}_{j,m})$$

where $V^{\text{out}}_{i,m}$ and $\hat{p}_{i,j,m}$ is defined in Equation 3.2 and 3.3, respectively. d is the measure of geographic distance between two coordinates, and $\text{home}_{i,m}$ represents the home coordinates for user i in month m.

doi:10.6342/NTU202502893



Appendix A — Event-Centered Trends Across Outcomes and Treatments demonstrates time series for two groups of features, showing the mean for the treated group against that of the untreated group, with all data points subtracted from the mean in the reference period. This allows us to better examine the outcome trends in both pre-treatment and post-treatment periods.

3.6.2 Human Mobility Features

Numerous studies have used CDRs to analyze human mobility patterns. We largely follow the literature in constructing mobility features but determine the weight of a telecom base station for a phone user by temporal size (see Equation 3.1) rather than call count, which corrects distortion caused by load sharing mechanisms. We use notations similar to those in Section 3.2 to avoid redundant redefinition.

There are two key differences. First, we now use CDRs from weekdays, including both daytime and nighttime records. Second, we add a monthly dimension, constructing features on a monthly basis. For example, $B_{i,m} \subset B$ denotes the collection of telecom base stations that handled calls for user i in month m, $R_{i,m,b} \subset R$ represents call records that are related to user i in month m and associated with base station $b \in B_{i,m}$, and $T_{i,m,b} \subset T$ denotes timestamps connected to $R_{i,m,b}$.

Human mobility features are derived purely from CDRs, with the locations of telecom base stations serving as proxies for visited places. Therefore, all geographic information is represented by a location matrix $L^{\text{geo}}_{i,m}$ defined as follows.

Definition 3.6.4 (Geographic Location Matrix)**.** A geographic location matrix $L^{\text{geo}}_{i,m} \in \mathbb{R}^{|B_{i,m}|\times 2}$ contains all visited coordinates for user i in month m, where the two columns

doi:10.6342/NTU202502893

record the longitude and latitude, respectively, of each telecom base station $b \in B_{i,m}$. Besides, we denote $(l^{\text{geo}}_{i,m})_b \in \mathbb{R}^2$ as the row of $L^{\text{geo}}_{i,m}$ that corresponds to telecom base station $b \in B_{i,m}$, which contains the longitude and latitude information of that telecom base station.

Mobility features often lay on the foundation of empirical probability distribution over the visited locations (rows of the location matrix). As mentioned, we consider a new weighting scheme where the weight for each $b \in B_{i,m}$ is defined as the temporal size $\mu(T_{i,m,b})$, and therefore, the corresponding empirical probability is given by:

$$\hat{p}^{\mu}_{i,m,b} = \frac{\mu(T_{i,m,b})}{\sum_{b' \in B_{i,m}} \mu(T_{i,m,b'})} \tag{3.4}$$

in contrast to the traditional approach:

$$\hat{p}^{T}_{i,m,b} = \frac{|T_{i,m,b}|}{\sum_{b' \in B_{i,m}} |T_{i,m,b'}|}, \tag{3.5}$$

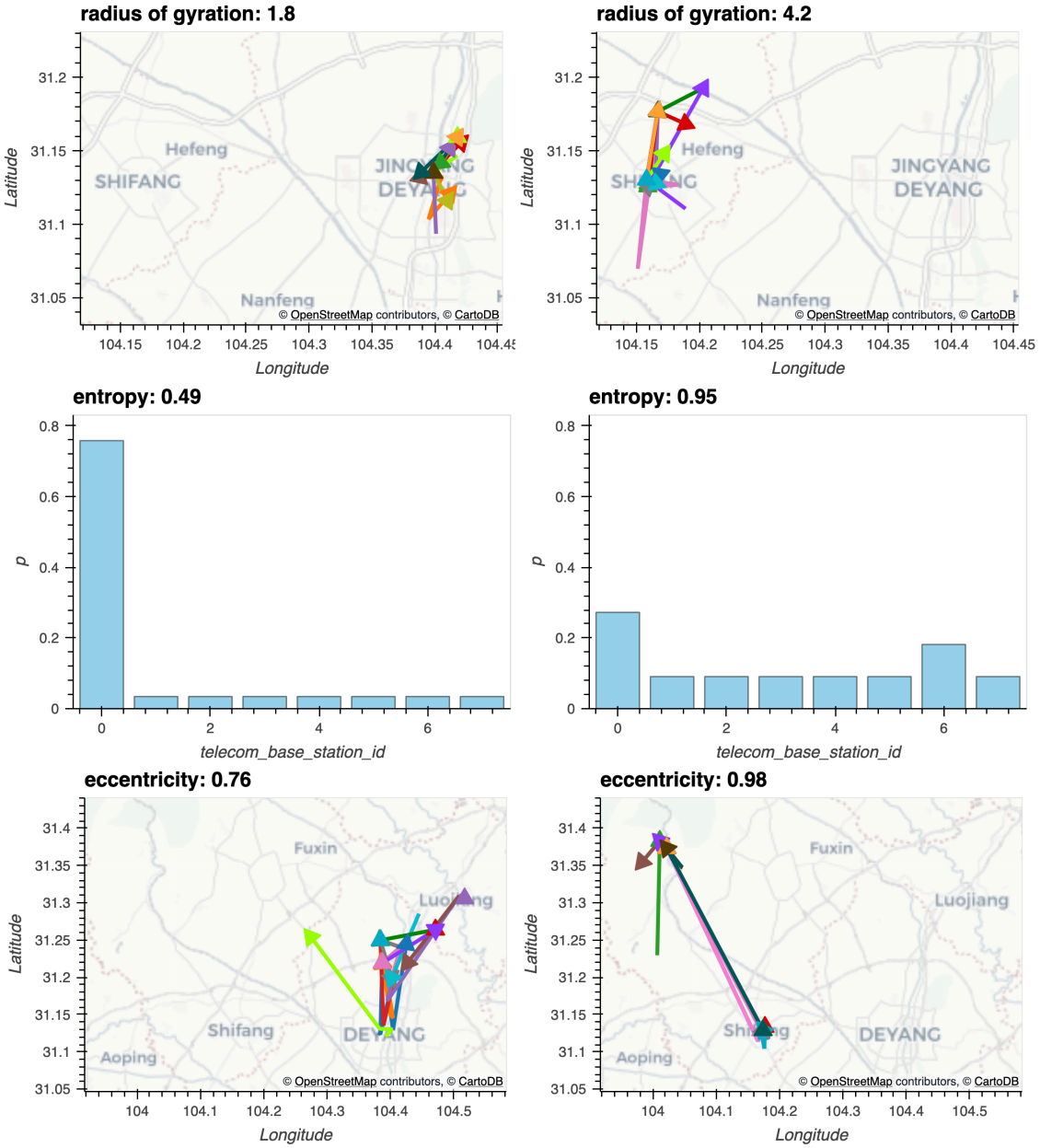
which is based on the count of call events.

Note that mobility features constructed using the temporal-size-based sample probability $\hat{p}^{\mu}_{i,m,b}$ will be named with the prefix "temporal-size-weighted," while those that depend on count-based sample probability $\hat{p}^{T}_{i,m,b}$ will be named with the prefix "count-weighted." For instance, count-weighted eccentricity versus temporal-size-weighted eccentricity.

In the following text, we will operate on the empirical probability vector $\hat{p}_{i,m} \in \mathbb{R}^{|B_{i,m}|}$, where each entry $\hat{p}_{i,m,b}$ corresponds to the empirical probability of base station $b \in B_{i,m}$ and can be computed using either the temporal size approach ($\hat{p}^{\mu}_{i,m,b}$) or the traditional call count approach ($\hat{p}^{T}_{i,m,b}$).

doi:10.6342/NTU202502893

Figure 3.4: Comparison of Mobility Feature Values



Notes: For radius of gyration and eccentricity, we visualize daily movement trajectories of four phone users throughout August 2013. Regarding entropy, we examine the distribution of telecom base station usage for two phone users during the same period, where the y-axis (p) represents the proportion of total days on which each base station was activated, calculated as the number of days a station served calls divided by the total number of such days across all stations.

Geographic distance (geodesic distance) on the Earth's surface differs from Euclidean distance calculated directly from longitude-latitude coordinates, as the former accounts for spherical geometry while the latter assumes a flat coordinate space. To make spatial

doi:10.6342/NTU202502893

statistical analysis meaningful, we employ the Azimuthal Equidistant (AEQD) projection[2] centered at the geographic mean $(r_c)_{i,m}^{\text{geo}}$ of locations, which is given by:

$$(r_c)_{i,m}^{\text{geo}} = (L_{i,m}^{\text{geo}})' \hat{p}_{i,m} \in \mathbb{R}^2. \tag{3.6}$$

This projection makes variance-covariance matrices computed on the projected coordinates geographically interpretable, as they accurately reflect the spatial spread of locations around the centroid. Besides, we denote $L_{i,m}$, $(l_{i,m})_b$, and $(r_c)_{i,m}$ as the projected location matrix, coordinates of a telecom base station b, and the centroid, respectively.

After completing the preliminary setup, we can now construct mobility features that aim to measure: (i) how large the activity area is (radius of gyration), (ii) how unpredictable spatial movement patterns are (movement entropy), and (iii) to what extent the activity area spreads in an elliptical shape (eccentricity), which indicates whether a user visits locations primarily along a fixed direction. See Figure 3.4 for a comprehensive understanding of what these metrics actually mean.

Definition 3.6.5 (Radius of Gyration)**.** The radius of gyration $(r_g)_{i,m}$ for user i in month m is defined by:

$$(r_g)_{i,m} = \sqrt{\sum_{b \in B_{i,m}} \hat{p}_{i,m,b} \cdot \|(l_{i,m})_b - (r_c)_{i,m}\|^2}.$$

Notably, it takes the form of a root-mean-square distance, as it's borrowed from physics, and the intuition is that we are thinking of users as orbiting around the center of mass $(r_c)_{i,m}$.

[2]The projection is obtained through the Python package pyproj. We choose AEQD over UTM to ensure that the Euclidean distance from any location contained in a location matrix to the centroid equals the geodesic distance. While UTM provides approximately correct Euclidean distances for all locations, we only require accuracy for distances to the centroid. Additionally, UTM zones are limited to 6° of longitude, but Sichuan spans approximately 11.6°, which would require multiple zones and result in non-uniform projections across users. Note that with AEQD using individual projection centers, all users' results are expressed in the same units (kilometers), ensuring comparability.

doi:10.6342/NTU202502893

Definition 3.6.6 (Movement Entropy)**.** The movement entropy $me_{i,m}$, where me stands for movement entropy, for user i in month m is defined by:

$$me_{i,m} = \frac{-\sum_{b \in B_{i,m}} \hat{p}_{i,m,b} \log(\hat{p}_{i,m,b})}{\log(|B_{i,m}|)}.$$

Definition 3.6.7 (Eccentricity)**.** Given $\hat{\Sigma}_{i,m}$ is the sample variance-covariance matrix of $L_{i,m}$, eccentricity $ecc_{i,m}$, where ecc stands for eccentricity, for user i in month m is defined as:

$$ecc_{i,m} = \sqrt{1 - \left(\frac{\lambda_{i,m,2}}{\lambda_{i,m,1}}\right)^2}$$

where $\lambda_{i,m,1}$ is the major eigenvalue of $\hat{\Sigma}_{i,m}$ and $\lambda_{i,m,2}$ is the minor one.

The sample variance-covariance matrix $\hat{\Sigma}_{i,m}$ is constructed in two steps. First, compute the demeaned location matrix $\tilde{L}_{i,m} = L_{i,m} - \mathbf{1}\hat{\mu}'_{i,m}$ where $\mathbf{1} \in \mathbb{R}^{|B_{i,m}|}$ is a vector of ones and $\hat{\mu}_{i,m}$ is the sample mean, which is equivalent to $(r_c)_{i,m}$. Then, the sample variance-covariance matrix is defined by:

$$\hat{\Sigma}_{i,m} = \kappa_{i,m}(\tilde{L}'_{i,m}\text{Diag}(\hat{p}_{i,m})\tilde{L}_{i,m})$$

where $\text{Diag}(\hat{p}_{i,m})$ is the diagonal matrix with the probability weights $\hat{p}_{i,m}$ on its diagonal, and $\kappa_{i,m}$ is the bias-correction factor, where

$$\kappa_{i,m} = \begin{cases} \dfrac{\sum_{b \in B_{i,m}} \mu(T_{i,m,b})}{\left(\sum_{b \in B_{i,m}} \mu(T_{i,m,b})\right) - 1} & \text{if } \hat{p}_{i,m,b} = \hat{p}^{\mu}_{i,m,b}, \\ \dfrac{\sum_{b \in B_{i,m}} |T_{i,m,b}|}{\left(\sum_{b \in B_{i,m}} |T_{i,m,b}|\right) - 1} & \text{otherwise.} \end{cases}$$

doi:10.6342/NTU202502893

3.7 Empirical Strategy

Appendix B — Preliminary of DiD Estimator has a brief introduction on the canonical 2×2 DiD design, which involves exactly two groups and two time periods. To examine treatment effect dynamics over multiple periods, we generalize the double comparison approach by selecting a baseline reference period—which corresponds to $t-1$ in the 2×2 DiD case—and then applying the DiD methodology to estimate treatment effects for each subsequent period relative to this reference point. DiD with multiple periods is also called event study, and we use these terms interchangeably. Conventionally, practitioners will utilize the two-way fixed effects specification to facilitate the estimation of average treatment effect on the treated in each period. Furthermore, the canonical 2×2 setup assumes static treatment timing, where all treated units receive treatment simultaneously. However, our empirical setting are different in the sense that treatment units become treated at different periods. This variation in treatment timing introduces additional complexity, and several econometrics tools have proposed to address this issue.

In the Definition 3.4.1 and 3.5.1, we include treatment units with staggered treatment adoption to avoid contemporaneous confounders, leading to a robust estimation. However, the design of DiD with multiple periods and staggered treatment adoption, through two-way fixed effect specification may be biased due to the forbidden comparison (this term also used by Roth et al. (2023) and De Chaisemartin and d' Haultfoeuille (2023)). That is, comparing the latter treated units to the early treated units, in the context of time-variant treatment effects (Goodman-Bacon (2021), Sun and Abraham (2021), Baker, Larcker and Wang (2022)). Through inspecting how mobility and mobile communication features evolve after residential shift (see Figure from A.1 to A.6) and smartphone adoption

doi:10.6342/NTU202502893

(see Figure from A.7 to A.12), we can observe the non-static treatment effect dynamics, resulting in the inapplicability of two-way fixed effects models.

Consequently, we follow the estimation approach proposed by Callaway and Sant' Anna (2021), which estimates the ATT dynamics for each treatment-timing cohort independently with the valid control group, thereby avoiding the issue of forbidden comparison. They name the ATT specific to a period and treatment-timing cohort as group-time ATT where the group is defined as the treatment-timing cohort. Therefore, in the following text, "groups" refers to treatment-timing cohorts while "cohorts" aside from groups, the control units are included as a single cohort.

Aside from Callaway and Sant' Anna (2021), many other econometric tools have been proposed to solve the issue as to the forbidden comparison, and Roth et al. (2023) and De Chaisemartin and d' Haultfoeuille (2023) both provide comprehensive introduction and briefly summarize the differences across various approaches. We choose Callaway and Sant' Anna (2021) over other alternatives as it's considered to be more flexible in (i) the selection of valid control units, (ii) the aggregation of estimated treatment effects across cohorts and periods, and (iii) the assumption on parallel trends.

As mentioned, the key point of dealing with the time-variant treatment effects with staggered treatment adoption is to employ valid control units. Callaway and Sant' Anna (2021) is flexible in the selection of valid control units in the sense that they allow either the never-treated units or the not-yet-treated units to be control units depending on the practitioners' need. However, others' control units (e.g., Sun and Abraham (2021) and Borusyak, Jaravel and Spiess (2024)) encompass of both never-treated and not-yet treated units, and don't have the freedom to choose either of them. Despite flexible, Callaway

doi:10.6342/NTU202502893

and Sant' Anna (2021) suggest employing the early-treated samples as the comparison only if the never-treated units are unavailable or limited in size. In our situation where only a small fraction of samples decides to change their homes or start using smartphones, never-treated samples are quite suitable as the comparison.

Callaway and Sant' Anna (2021) aim at estimating ATT for each group and period, and therefore, researcher has the full customizability to aggregate them across groups and periods to obtain a summary ATT. Additionally, they incorporate covariates on which the parallel trends assumption (hereafter PTA) conditions, which should be reasonable when the unconditional PTA is violated because groups differ in observable characteristics that affect outcome trends. Nevertheless, conditional PTA introduces additional layer of complexity for estimation, so we will first start from the unconditional version for estimation, and if we clearly see the patterns of violation, we will move to incorporate the pre-treatment covariates, adopting the conditional PTA. Furthermore, the anticipation is allowed, which is particularly useful when discussing the impacts of residential shift as mobility or mobile communication features might start to change prior to the actual relocation timing.

3.8 Group-Time ATT

Before introducing the group-time ATT, we need to set up the potential outcome framework first to let it have a clear definition. Since we include treatment units with four different treatment timings ($\{4, 5, 6, 7\} \subset \mathcal{M}$), for each user i, in each month $m \in \mathcal{M}$, there will be 5 potential outcomes: $Y_{i,m}(\infty)$, $Y_{i,m}(4)$, $Y_{i,m}(5)$, $Y_{i,m}(6)$, and $Y_{i,m}(7)$. $Y_{i,m}(\infty)$ is the potential outcome in month m for user i if i has never received the treat-

doi:10.6342/NTU202502893

ment throughout the sample period. $Y_{i,m}(g)$ where $g \in \{4,5,6,7\} \subset \mathcal{M}$ is the potential outcome in month m for user i if i is treated in month g (i.e., $i \in \mathcal{G}_g$). For each user i and month m, only 1 potential outcome can be realized, becoming observable. Hence, the connection between potential outcomes and the observed outcome $Y_{i,m}$ is established (Callaway and Sant' Anna (2021), Sun and Abraham (2021)) as follows:

$$Y_{i,m} = Y_{i,m}(\infty) + \Sigma_{g=4}^{7}(Y_{i,m}(g) - Y_{i,m}(\infty)) \cdot G_{i,g} \tag{3.7}$$

where $G_{i,g}$ is a binary variable defined as:

$$G_{i,g} = \begin{cases} 1, & \text{if } i \in \mathcal{G}_g, \text{ i.e., } i \text{ receives the treatment in month } g \\ 0, & \text{otherwise.} \end{cases}$$

Besides, we define an additional binary variable for indicating whether i is in the never-treated cohort g_∞:

$$G_{i,\infty} = \begin{cases} 1, & \text{if } i \in \mathcal{G}_\infty \\ 0, & \text{otherwise.} \end{cases}$$

Anticipation is a critical topic when estimating the ATT as it determines which pre-treatment period is considered as the reference to correctly assess treatment effects. Let δ be the number of anticipation months, and for cohort $\mathcal{G}_g$, $g - \delta$ becomes the cutoff value (note that if $\delta = 0$, the cutoff value is g, which returns to the conventional setting) and treatment effect dynamics are inspected for each month $m \geq g - \delta$, as treatment effects are expected to let $Y_{i,m}(g)$ deviate from $Y_{i,m}(\infty)$ starting at any timing m after $g - \delta$. This also implies:

$$Y_{i,m}(g) = Y_{i,m}(\infty) \tag{3.8}$$

doi:10.6342/NTU202502893

for all $i \in \mathcal{G}_g$ and $m < g - \delta$ (modified from the Assumption 5 in Roth et al. (2023)). If we jointly consider the Equation 3.7 and 3.8, we can further derive:

$$Y_{i,m} = Y_{i,m}(g) = Y_{i,m}(\infty) \tag{3.9}$$

for all $i \in \mathcal{G}_g$ and $m < g-\delta$. With these potential outcome notations, it's time to define the group-time ATT, which is ATT for a specific treatment-timing group in a specific priod.

Definition 3.8.1 (Group-Time ATT)**.** The ATT specific to treatment-timing cohort $\mathcal{G}_g$ in month $m \in \mathcal{M}$ is defined as:

$$ATT(g, m) = \mathbb{E}[Y_{i,m}(g) - Y_{i,m}(\infty) \mid G_{i,g} = 1],$$

which is the expected difference between potential outcomes in month m for cohort $\mathcal{G}_g$.

As user i belongs to the cohort $\mathcal{G}_g$, by equation 3.7, we know $Y_{i,m}(g)$ is observable and equivalent to $Y_{i,m}$. Therefore, the group-time ATT can be rewritten as:

$$\begin{aligned} ATT(g, m) &= \mathbb{E}[Y_{i,m} - Y_{i,m}(\infty) \mid G_{i,g} = 1] \\ &= \mathbb{E}[Y_{i,m} \mid G_{i,g} = 1] - \mathbb{E}[Y_{i,m}(\infty) \mid G_{i,g} = 1], \end{aligned} \tag{3.10}$$

which is the difference between the expected observed outcome $Y_{i,m}$ with respect to cohort $\mathcal{G}_g$ and the expected potential outcome in month m that would occur if they have never been exposed to the treatment.

Note that notations in Definition 3.8.1 are a little bit different from Callaway and Sant' Anna (2021) in the way that they denote the period as t and causal parameter as $ATT(g, t)$ while we use m to emphasize the monthly periods. Besides, we also replace

doi:10.6342/NTU202502893

the 0 with ∞ used by Sun and Abraham (2021) and Borusyak, Jaravel and Spiess (2024) to align the meaning of never-treated units, which become treated at some point in the infinite future.

Given the anticipation months δ, and considering a treatment-timing cohort $\mathcal{G}_g$, for each $m \geq g - \delta$, $Y_{i,m}(\infty)$ is unobservable. Therefore, we need an assumption to make $ATT(g, m)$ identifiable for month $m \geq g - \delta$, and that's where PTA comes into play.

Definition 3.8.2 (Parallel Trend Assumption)**.** Given the number of anticipation months δ, for each $\mathcal{G}_g$ where $g \in \{4, 5, 6, 7\}$ and $m \geq g - \delta$, the following equality is assumed:

$$\mathbb{E}[Y_{i,m}(\infty) - Y_{i,m-1}(\infty) \mid G_{i,g} = 1] = \mathbb{E}[Y_{i,m}(\infty) - Y_{i,m-1}(\infty) \mid G_{i,\infty} = 1]. \quad (3.11)$$

That is, for all group $\mathcal{G}_g$ and month $m \geq g - \delta$, the trend in the counterfactual untreated outcome measured between month m and $m - 1$ should be expectedly identical between group $\mathcal{G}_g$ and the never-treated cohort $\mathcal{G}_\infty$ for all $\mathcal{G}_g$. It's an assumption as objects in the equality are potential outcomes $Y_{i,m}(\infty)$, which are clearly unobservable in the post-treatment periods for all $\mathcal{G}_g$. As mentioned, we will impose the unconditional PTA first, and switch to the conditional version if needed.

The parallel trend assumption stated in Definition 3.8.2 is employed in the canonical 2×2 DiD design, and it's necessary to extend it for the multiple periods case, which is given by:

$$\mathbb{E}[Y_{i,m}(\infty) - Y_{i,g-\delta-1}(\infty) \mid G_{i,g} = 1] = \mathbb{E}[Y_{i,m}(\infty) - Y_{i,g-\delta-1}(\infty) \mid G_{i,\infty} = 1]. \quad (3.12)$$

Equation 3.12 can be obtained from Equation 3.11 by adding all parallel trend equality in post-treatment periods and by the end, $Y_{i,g-\delta-1}(\infty)$ appears, which is observable since

doi:10.6342/NTU202502893

$Y_{i,g-\delta-1}(\infty) = Y_{i,g-\delta-1}$. As mentioned in Roth et al. (2023), for the ATT in the longer periods after treatment to be identifiable, a stronger assumption needs to be imposed.

With equation 3.12, $\mathbb{E}[Y_{i,m}(\infty) \mid G_{i,g} = 1]$ becomes identifiable:

$$\begin{aligned}
\mathbb{E}[Y_{i,m}(\infty) \mid G_{i,g} = 1] &= \mathbb{E}[Y_{i,g-\delta-1}(\infty) \mid G_{i,g} = 1] \\
&\quad + \mathbb{E}[Y_{i,m}(\infty) \mid G_{i,g} = 1] - \mathbb{E}[Y_{i,g-\delta-1}(\infty) \mid G_{i,g} = 1] \\
&= \mathbb{E}[Y_{i,g-\delta-1} \mid G_{i,g} = 1] \\
&\quad + \underbrace{\mathbb{E}[Y_{i,m}(\infty) \mid G_{i,\infty} = 1] - \mathbb{E}[Y_{i,g-\delta-1}(\infty) \mid G_{i,\infty} = 1]}_{\text{due to parallel trends}} \\
&= \mathbb{E}[Y_{i,g-\delta-1} \mid G_{i,g} = 1] + \mathbb{E}[Y_{i,m} - Y_{i,g-\delta-1} \mid G_{i,\infty} = 1].
\end{aligned} \tag{3.13}$$

Then, group-time ATT can be derived by:

$$\begin{aligned}
ATT(g,m) &= \mathbb{E}[Y_{i,m} \mid G_{i,g} = 1] - \mathbb{E}[Y_{i,m}(\infty) \mid G_{i,g} = 1] \\
&= \mathbb{E}[Y_{i,m} \mid G_{i,g} = 1] \\
&\quad - \mathbb{E}[Y_{i,g-\delta-1} \mid G_{i,g} = 1] - \mathbb{E}[Y_{i,m} - Y_{i,g-\delta-1} \mid G_{i,\infty} = 1] \\
&= \mathbb{E}[Y_{i,m} - Y_{i,g-\delta-1} \mid G_{i,g} = 1] - \mathbb{E}[Y_{i,m} - Y_{i,g-\delta-1} \mid G_{i,\infty} = 1].
\end{aligned} \tag{3.14}$$

By Equation 3.14, we can simply estimate $ATT(g,m)$ through its sample analogue. The simplicity credits to the unconditional PTA, and for how to incorporate conditional PTA, please refer to Callaway and Sant' Anna (2021). Note that statistical inference relies on the bootstrap procedure for standard error estimation.

It is important to note that PTA is extremely crucial for identification, as the ATT becomes unidentifiable without it. This raises the question whether there exists a method

doi:10.6342/NTU202502893

to test this assumption. Typically, the assumption's plausibility is evaluated in the pre-treatment periods in an event study design. As in pre-treatment periods, the potential outcome of the never-treated $Y_{i,m}(\infty)$ is observable for all user i, and if the PTA holds in the pre-treatment periods, it might be relatively reasonable to claim that the parallel trend might also hold in post-treatment periods.

That's the reason why we plot the treated group's trajectories of mean outcomes centered at the reference period $g - \delta - 1$, along with those of the control group. In all figures in Appendix A — Event-Centered Trends Across Outcomes and Treatments, we can see that the treated group's re-centered mean outcomes' trajectories highly overlaps with those associated to the control group's trajectory in the pre-treatment periods. However, exceptions occur in specific treatment-timing cohorts for certain outcomes, raising some concerns about the plausibility of the PTA.

Moreover, given that the treatment is not confounded, and anticipation effects are correctly accounted for, if PTA holds, the estimated ATT in pre-treatment periods should be statistically insignificant. The intuition is that the treatment shouldn't take effects during pre-treatment periods and under PTA, treated units' outcome changes should be identical to the comparison group's changes over the same periods, yielding insignificant ATT estimation. Therefore, by examining the estimation results in pre-treatment periods, we can assess whether the PTA holds during the pre-treatment periods.

doi:10.6342/NTU202502893

After estimating the group-time ATT, we can recover the traditional event study through the following aggregation scheme provided by Callaway and Sant' Anna (2021), which demonstrates the ATT after e month of the treatment:

$$\theta(e) = \sum_{g=4}^{7} \mathbb{P}(G_{i,g} = 1) ATT(g, g+e) 1[g+e <= 10] \tag{3.15}$$

where $\mathbb{P}(G_{i,g} = 1)$ is the group size of cohort $\mathcal{G}_g$ and $1[g+e <= 10]$ is an indicator function that equals to 1 if $g+e <= 10$ and 0 otherwise. Note that 10, representing May 2014, is the maximum month in our sample period. Moreover, we restrict e to be from -3 to 3 to let the difference of $\theta(e)$ be the correct interpretation of treatment effect dynamics. The Intuition is within this event time interval, the share of group size of each treatment-timing cohort is fixed. For example, when $e = 4$, the outcomes of cohort $\mathcal{G}_7$ are missing and when $e = -4$, the outcomes of cohort $\mathcal{G}_4$ are missing. For more details, please refer to the Equation 3.5 in Callaway and Sant' Anna (2021).

doi:10.6342/NTU202502893

Chapter 4 Results

4.1 Outcomes of Interest

Although we can construct the mobile communication network features based on two directions of mobile communication—incoming and outgoing—we will focus on the outgoing direction as the variation in outgoing-based features may be more sensitive to the factors directly linked to the treated units themselves. Moreover, through observing Figures in Appendix A — Event-Centered Trends Across Outcomes and Treatments, where various outcomes' paths during the whole sample period are plotted, we can see that the evolution patterns typically don't have dramatical differences between outgoing and incoming communication. Therefore, when discussing the treatment effects of both residential shift and smartphone adoption, we will consider only the outgoing-based mobile communication features. For mobility features, although we propose a new weighting scheme to address the issue of spurious importance of telecom base stations for phone users, arising from the load sharing mechanism, it seems that mobility features' differences between two construction methods based on distinct weighting schemes are subtle (see Appendix A — Event-Centered Trends Across Outcomes and Treatments). Hence, we will adopt our proposed weighting scheme based on the concept of temporal size to compute the mobility features. All in all, outcomes consists of two groups: mobility fea-

doi:10.6342/NTU202502893

tures and mobile communication network features, and there are three outcomes in each group. The total of six outcomes focuses on capturing different aspects of human's mobility and mobile communication behavior.

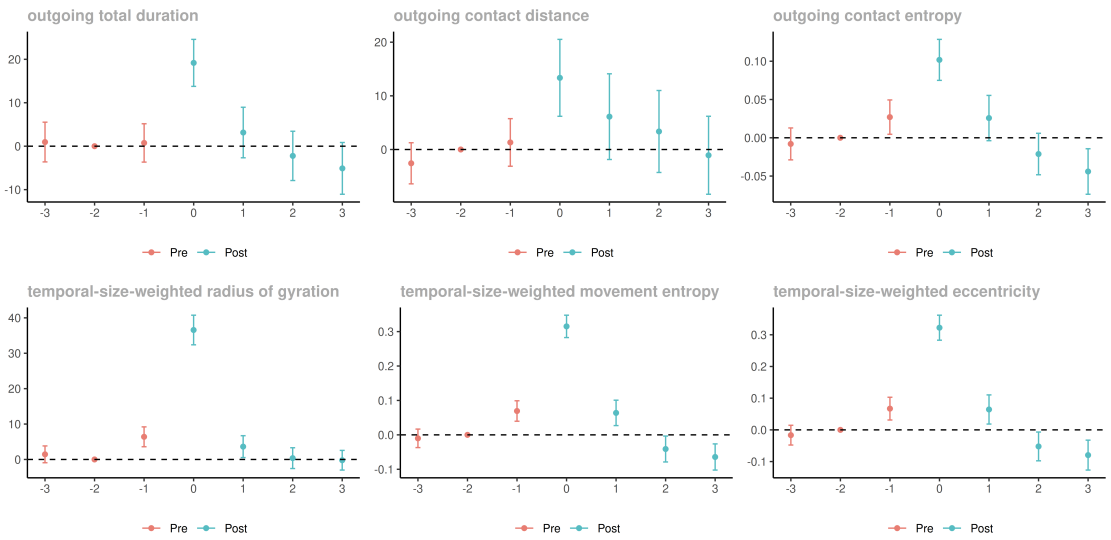
4.2 Residential Shifts

In the current and next section, we are going to analyze the estimation results of ATT with the delicately determined anticipation parameter (see Selection of Anticipation Parameter) and the assumption of unconditional PTA. To understand the magnitudes of the estimated effects, we will naively employ the mean outcome of the control group in September 2013 as the baseline for the context of residential shifts, while for smartphone adoption, the baseline will be that of October 2013. The confidence interval is constructed based on the 95% significance level. We first analyze heterogeneous treatment effects across time periods. Specifically, we will approach with a hierarchical manner starting from the discussion of static versus time-variant treatment effects. The static treatment effect means the outcome is persistently shifted without fading back to the baseline level. In such situation, attribution includes upward or downward shift. For the time-variant one, we can characterize it to be either transient or smoothly decaying. A transient treatment effect over time represents an effect that instantly bounces back and forth to the baseline, while smoothly decaying is the other case, where the effect gradually fades away. Then, we can discuss the heterogeneous effect across treatment-timing groups by inspecting whether only few groups exhibit distinct patterns.

At the very first glance, we can see that the residential shift brings about time-variant effects on both mobile communication network (upper panel) and mobility features (lower

doi:10.6342/NTU202502893

Figure 4.1: Aggregated Event Study of Residential Shifts



panel), and effects fall back to or approach to the baseline level one month after relocations begin.

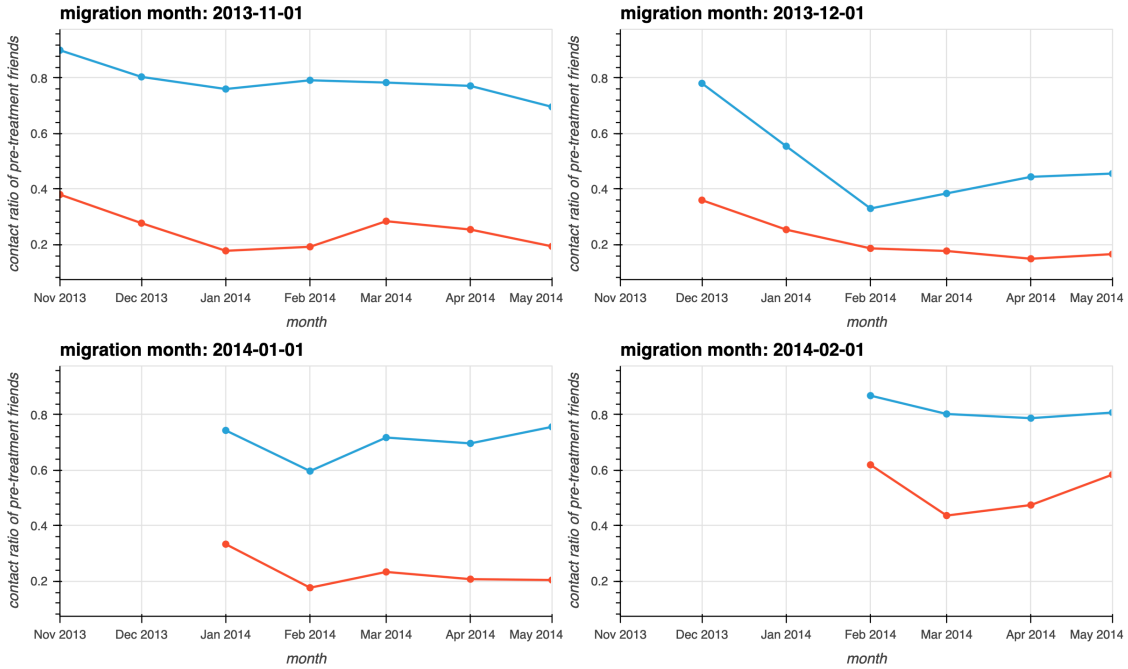
For the mobile communication behavior, anticipation for residential relocation is limited and observable only in outgoing contact entropy. The pre-treatment shift in outgoing contact entropy might be result from notifying forthcoming relocation with friends that are less frequently contacted. However, from Figure D.13, we can see that this phenomenon only emerges in the treated units who migrate in February 2014 (group 7).

Intuitively, residential relocations expose mobile phone users to unfamiliar environments, resulting in a temporary surge in total call duration during the month of relocation. Relative to the baseline mean of 89.7 minutes, migrants experience a 19 minute (21%) increase in outgoing call duration, and a 13 km (189%) increase in outgoing contact distance relative to the baseline mean of 6.9 km as they seek to connect with geographically distant social connections. Moreover, during this same period, migrants engage in more diversified social interactions, with entropy increasing by 0.10 units (a 17% increase from the baseline of 0.60), likely due to the formation of new social connections at the destination or

doi:10.6342/NTU202502893

contact with those with whom they engaged less frequently prior to the relocation Nevertheless, following the completion of relocation, migrants tend to spend less time on mobile communication with less diversified interactions. By the third month post-relocation, total call duration decreases by 5.1 minutes (a 6% reduction from the 89.7-minute baseline) and contact entropy falls by 0.04 units (a 7% decrease from the baseline of 0.60), indicating a return toward more concentrated communication patterns.

Figure 4.2: Contact Ratio of Pre-Treatment Friends in Post-Treatment Periods



Notes: The blue line is the contact ratio of pre-treatment friends while the red line additionally require those in the same origin prefecture.

In Figure 4.2, we analyze the composition of migrants' contacts in post relocation periods. We define a new variable called contact ratio, and the contact ratio of a user in a given month measures the proportion of the users' call duration with friends with whom they are already connected versus all friends contacted in the corresponding month. In the above plot, we present the monthly contact ratio series obtained by averaging across all users in each given month. We also separately plot the contact ratios for migrants who relocate in different months though they seem to behave very similarly.

doi:10.6342/NTU202502893

We can confirm that, in the early periods of post-migration, migrants tend to connect with those old friends with whom are already connected prior to the migration events. Typically, in the same months of changing residential places, around 80% of contacts are pre-treatment friends, but they are not necessarily in the same prefectures from which migrants come. Specifically, roughly 40% of these pre-treatment friends reside in the migrants' origin prefecture. These facts support the hypothesis that the sudden surge in mobile communication during the period of migration is due to connections with existing friendships. Moreover, the drastic increase in contact distance also suggests that migrants move away from their friends rather than toward them.

For mobility patterns, the evidence of anticipation behavior is even more solidified, with moderate upward shift. The spatial coverage of movement activity measured through the metric of radius of gyration substantially increases contemporaneously with the completion of residential relocation, expanding by 36.6 km (a 412% increase from the 8.9 km baseline) while the effect quickly fades back in the following months. This substantial increase includes the component of traveling distance between the origin and destination locations during the relocation process. Movement entropy and eccentricity evolve in an extreme pattern where migrants instantly transit from highly unpredictable spatial appearances with spatial stretching along a fixed direction to predictable patterns with roughly circular spatial distribution. This is characterized by a dramatic shift from 0.32 above the entropy baseline of 0.67 (a 47% increase) and 0.32 above the eccentricity baseline of 0.82 (a 39% increase) during relocation month, to -0.06 (9% below baseline) and -0.08 (10% below baseline) respectively by the third month post-relocation. The intuition is that when moving to a new environment, a migrant might be actively exploring but have limited knowledge, resulting in exploration along a fixed axis. As time passes, exploration

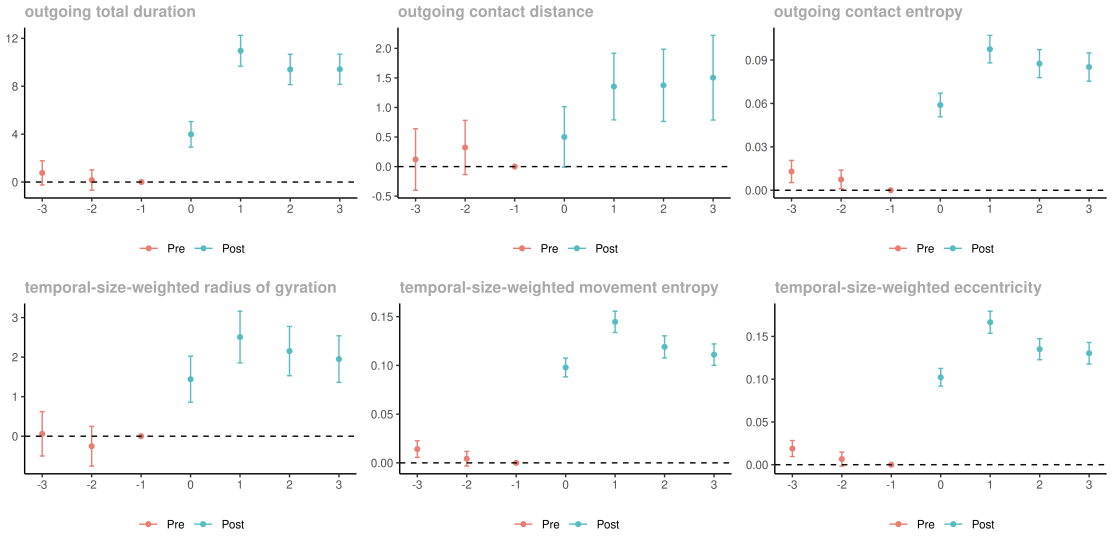
doi:10.6342/NTU202502893

patterns settle down, and they mainly visit a few places located in various directions, creating a more balanced and predictable spatial distribution.

Both groups of features seem to be continuously evolving to negative states. To better understand the evolving paths of various features across treatment-timing groups, please refer to Appendix D — Group-Specific Event Studies.

4.3 Smartphone Adoption

Figure 4.3: Aggregate Event Study of Smartphone Adoption



Unlike the residential shift, we expect no anticipation for smartphone adoption. We clearly see that the effects are nearly static over time and the shifts are positive, but the magnitude is relatively small compared to the residential shift. By inspecting ATT dynamics across different groups demonstrated in both Figure D.15 and D.16, we can confirm that there is no heterogeneous effect across groups.

Relative to the non-smartphone users' baseline in October 2013, smartphone adoption leads to immediate behavioral changes. The most notable influence is the increase

doi:10.6342/NTU202502893

in radius of gyration by 1.4 km (a 21% increase from the 6.8 km baseline). There are also modest shifts in movement entropy (0.10 units, 16% increase from the 0.62 baseline) and eccentricity (0.10 units, 13% increase from the 0.78 baseline), along with smaller increases in contact entropy of 0.06 units (10% increase from the 0.57 baseline) and call duration of 4.0 minutes (9% increase from the 44.8 minutes baseline).

The shifts in movement entropy and eccentricity likely reflect the integration of geographical technologies, such as online maps, facilitating exploration of unfamiliar places. Interestingly, this pattern is identical to that observed during residential relocations, where individuals moving to new environments exhibit increased movement entropy (exploring behavior) coupled with increased eccentricity (fixed directional preference). This finding additionally confirms the systematic "exploring pattern of individuals" where users do not tend to explore randomly but rather follow structured exploration along preferred directions, suggesting that smartphone adoption triggers similar exploratory behaviors as physical relocation to unfamiliar territories.

It seems counterintuitive that after upgrading to smartphone devices, users increase their call duration as they might have better access to the internet and switch to using mobile communication apps like WeChat. Our explanation is that our examination of effect dynamics is on a monthly basis, which is short-term, and these users originally used non-smartphones, which might suggest that they don't have a strong preference for the internet and tend to utilize phone calls to contact their friends. Therefore, after upgrading to smartphones, in the short run, they might still mainly rely on phone calls rather than the internet to interact with their friends. Besides, we believe that mobile infrastructure was not well implemented back in 2013 when 4G technology was not widely spread. Therefore, phone calls still served as a major tunnel for remote social interactions.

doi:10.6342/NTU202502893

Chapter 5 Discussion

5.1 Summary

Firstly, our methodological contributions is based on the incorporation of load sharing mechanisms. Specifically, we redesign the home location estimation approach, residential shift identification, and human mobility and mobile communication network features that are commonly used when utilizing CDRs to analyze these behaviors. Moreover, we present a new research topic that derives treatments from CDRs or user profile information and examines the dynamics of treatment effects on two common behaviors, going beyond conventional correlation studies utilizing CDRs. Finally, investigations into the impacts of residential shifts and smartphone adoption on mobility and communication behavior provide rich policy insights. For instance, our findings suggest policymakers should increase awareness of telecommunication infrastructure and urban transportation planning during large-scale population displacement and mobile technology transitions.

5.2 Limitations

A notable limitation in utilizing CDRs for home/work location inference and mobility pattern analysis is the inherent sampling bias present in the data. CDRs only capture spatial

doi:10.6342/NTU202502893

information at discrete moments when individuals initiate or receive telecommunications, thereby providing an incomplete representation of their complete spatial-temporal trajectory. This selective sampling characteristic potentially introduces bias into our inference methodologies. However, according to empirical investigations by Ranjan et al. (2012) and Zhao et al. (2016), while movement entropy estimates may exhibit either upward or downward bias depending on context, metrics such as radius of gyration and home/work location inferences demonstrate robust reliability. By extension, we posit that eccentricity measurements remain relatively unbiased, as they share fundamental characteristics with radius of gyration—specifically, both metrics aim to capture the geographical shape defined by the visited locations. Furthermore, the mobile communication network features we derive are specifically designed to quantify distinct contact behaviors, rendering them methodologically appropriate for our analytical framework despite the aforementioned sampling considerations.

Another limitation is that individuals, especially younger generations, have recently started engaging social media platforms more frequently to connect with friends and making fewer phone calls (Garrett et al. (2023)), which may deteriorate the quality of utilizing CDRs for home/work location inference and mobility pattern analysis. While this limitation does not affect our study since our sample period spans from 2013 to 2014, when social media platforms were not yet widespread, it raises questions about the validity of utilizing CDRs for more contemporary mobility and communication studies. In fact, there is a growing body of research attempting to leverage geolocated posts on various social media platforms, such as Facebook (Sahai and Bailey (2022)), Twitter (Zagheni et al. (2014), Hawelka et al. (2014), Jurdak et al. (2015), Luo et al. (2016)), and Weibo(Cui, Xie and Liu (2018), Ebrahimpour et al. (2020)), to study migration and human mobility

doi:10.6342/NTU202502893

patterns. Nevertheless, we still believe CDRs have their own advantages. First, in developing countries where network infrastructures are not well-developed, people still rely on phone calls to communicate with their family and friends. Second, people often use social media to share their travel experiences, resulting in important locations inferred from geolocated posts that potentially yield systematic biases. As stated in Armstrong et al. (2021), utilizing tweets to infer migration populations yields high misclassification rates. Finally, mobile communication is a more fundamental contact behavior, potentially leading to higher coverage of different age groups and reducing income bias due to unequal access to the internet, as Facebook users are often located in high-income regions in India (Sahai and Bailey (2022)).

5.3 Future Work

The main goal of this paper is to quantify the effects of residential shifts and smartphone adoption on behavioral features. We apply the estimation method proposed by Callaway and Sant'Anna (2021) and rely on the parallel trends assumption to identify treatment effects that are robust to heterogeneous treatment effects across time and treatment-timing groups. However, we acknowledge that we did not rigorously test for endogeneity, as both treatments are self-selected. For example, although we include multiple treatment-timing groups, our short-term sample periods cannot fully avoid contemporaneous factors at the annual level that affect treatment adoption and outcomes simultaneously. Future research should consider instrumental variable approaches to address these endogeneity concerns more rigorously.

doi:10.6342/NTU202502893

References

Armstrong, Caitrin, Ate Poorthuis, Matthew Zook, Derek Ruths, and Thomas Soehl. 2021. "Challenges when identifying migration from geo-located Twitter data." *EPJ Data Science*, 10(1): 1.

Ayesha, Buddhi, Bhagya Jeewanthi, Charith Chitraranjan, Amal Shehan Perera, and Amal S Kumarage. 2019. "User localization based on call detail record." 411–423, Springer.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics*, 144(2): 370–395.

Barwick, Panle Jia, Yanyan Liu, Eleonora Patacchini, and Qi Wu. 2023. "Information, Mobile Communication, and Referral Effects." *American Economic Review*, 113(5): 1170–1207.

Blumenstock, Joshua E. 2012. "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda." *Information Technology for Development*, 18(2): 107–125.

Blumenstock, Joshua E, Guanghua Chi, and Xu Tan. 2025. "Migration and the value of social networks." *Review of Economic Studies*, 92(1): 97–128.

doi:10.6342/NTU202502893

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting poverty and wealth from mobile phone metadata." *Science*, 350(6264): 1073–1076.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event-study designs: robust and efficient estimation." *Review of Economic Studies*, 91(6): 3253–3285.

Boustan, Leah Platt, Price V Fishback, and Shawn Kantor. 2010. "The effect of internal migration on local labor markets: American cities during the Great Depression." *Journal of Labor Economics*, 28(4): 719–746.

Bryan, Gharad, and Melanie Morten. 2019. "The aggregate productivity effects of internal migration: Evidence from Indonesia." *Journal of Political Economy*, 127(5): 2229–2268.

Büchel, Konstantin, Maximilian V Ehrlich, Diego Puga, and Elisabet Viladecans-Marsal. 2020. "Calling from the outside: The role of networks in residential mobility." *Journal of urban economics*, 119: 103277.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of econometrics*, 225(2): 200–230.

Chi, Guanghua, Fengyang Lin, Guangqing Chi, and Joshua Blumenstock. 2020. "A general approach to detecting migration events in digital trace data." *PloS one*, 15(10): e0239408.

Cho, Eunjoon, Seth A Myers, and Jure Leskovec. 2011. "Friendship and mobility: user movement in location-based social networks." 1082–1090.

doi:10.6342/NTU202502893

Cui, Yilan, Xing Xie, and Yi Liu. 2018. "Social media and mobility landscape: Uncovering spatial patterns of urban human mobility with multi source data." *Frontiers of Environmental Science & Engineering*, 12: 1–14.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2023. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." *The econometrics journal*, 26(3): C1–C30.

Dias, Viren, Lasantha Fernando, Yusen Lin, Vanessa Frias-Martinez, and Louiqa Raschid. 2022. "Framework to Study Migration Decisions Using Call Detail Record (CDR) Data." *IEEE Transactions on Computational Social Systems*, 10(5): 2725–2738.

Domínguez, Daniel Rodríguez, Rebeca P Díaz Redondo, Ana Fernández Vilas, and Mohamed Ben Khalifa. 2017. "Sensing the city with Instagram: Clustering geolocated data for outlier detection." *Expert systems with applications*, 78: 319–333.

Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. "Network diversity and economic development." *Science*, 328(5981): 1029–1031.

Ebrahimpour, Zeinab, Wanggen Wan, José Luis Velázquez García, Ofelia Cervantes, and Li Hou. 2020. "Analyzing social-geographic human mobility patterns using large-scale social media data." *ISPRS International Journal of Geo-Information*, 9(2): 125.

Espíndola, Aquino L, Jaylson J Silveira, and TJP Penna. 2006. "A Harris-Todaro agent-based model to rural-urban migration." *Brazilian journal of physics*, 36: 603–609.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. "A density-based algorithm for discovering clusters in large spatial databases with noise." Vol. 96, 226–231.

doi:10.6342/NTU202502893

Garrett, Shedrick L, Kaitlyn Burnell, Emma L Armstrong-Carter, Mitchell J Prinstein, and Eva H Telzer. 2023. "Linking video chatting, phone calling, text messaging, and social media with peers to adolescent connectedness." *Journal of Research on Adolescence*, 33(4): 1222–1234.

Gonzalez, Marta C, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. "Understanding individual human mobility patterns." *nature*, 453(7196): 779–782.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of econometrics*, 225(2): 254–277.

Hartigan, John A. 1975. *Clustering algorithms.* John Wiley & Sons, Inc.

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. "Geo-located Twitter as proxy for global mobility patterns." *Cartography and geographic information science*, 41(3): 260–271.

Hunt, Gary L, and Richard E Mueller. 2004. "North American migration: returns to skill, border effects, and mobility costs." *Review of Economics and Statistics*, 86(4): 988–1007.

Imbert, Clement, Marlon Seror, Yifan Zhang, and Yanos Zylberberg. 2022. "Migrants and firms: Evidence from china." *American Economic Review*, 112(6): 1885–1914.

Isaacman, Sibren, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. "Identifying important places in people' s lives from cellular network data." 133–151, Springer.

Jabbar, MA, and S Suharjito. 2020. "Fraud detection call detail record using machine learning in telecommunications company." *Adv. sci. technol. eng. syst. j*, 5: 63–69.

doi:10.6342/NTU202502893

Jurdak, Raja, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. "Understanding human mobility from Twitter." *PloS one*, 10(7): e0131469.

Karahoca, ADEM, and Ali Kara. 2006. "Comparing clustering techniques for telecom churn management." 27–29.

Lai, Shengjie, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Nick W Ruktanonchai, Alessandro Sorichetta, Jessica Steele, Tracey Li, Claire A Dooley, and Andrew J Tatem. 2019. "Exploring the use of mobile phone data for national migration statistics." *Palgrave communications*, 5(1): 1–10.

Luo, Feixiong, Guofeng Cao, Kevin Mulligan, and Xiang Li. 2016. "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago." *Applied Geography*, 70: 11–25.

Luo, Xusen, Yunyao Zhou, Yifu Yang, and Shuyun Wu. 2020. "Research on home and work locations based on mobile phone data." Vol. 1486, 052013, IOP Publishing.

Márquez-Barja, Johann, Carlos T Calafate, Juan-Carlos Cano, and Pietro Manzoni. 2011. "An overview of vertical handover techniques: Algorithms, protocols and tools." *Computer communications*, 34(8): 985–997.

Onnela, J-P, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. "Structure and tie strengths in mobile communication networks." *Proceedings of the national academy of sciences*, 104(18): 7332–7336.

doi:10.6342/NTU202502893

Pappalardo, Luca, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. "Returners and explorers dichotomy in human mobility." *Nature communications*, 6(1): 8166.

Pappalardo, Luca, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. 2016. "An analytical framework to nowcast well-being using mobile phone data." *International Journal of Data Science and Analytics*, 2: 75–92.

Phithakkitnukoon, Santi. 2022. "Inferring and Modeling Migration Flows Using Mobile Phone CDR Data." In *Urban Informatics Using Mobile Network Data: Travel Behavior Research Perspectives*. 75–101. Springer.

Phithakkitnukoon, Santi, Zbigniew Smoreda, and Patrick Olivier. 2012. "Socio-geography of human mobility: A study using longitudinal mobile phone data." *PloS one*, 7(6): e39253.

Ranjan, Gyan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. 2012. "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3): 33–44.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics*, 235(2): 2218–2244.

Sahai, Harshil, and Michael Bailey. 2022. "Social networks and spatial mobility: Evidence from Facebook in India." *arXiv preprint arXiv:2203.05595*.

Shi, Jieming, Nikos Mamoulis, Dingming Wu, and David W Cheung. 2014. "Density-based place clustering in geo-social networks." 99–110.

doi:10.6342/NTU202502893

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. "Limits of predictability in human mobility." *Science*, 327(5968): 1018–1021.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of econometrics*, 225(2): 175–199.

Tongsinoot, Lumpsum, and Veera Muangsin. 2017. "Exploring home and work locations in a city from mobile phone data." 123–129, IEEE.

Wang-Lu, Huaxin, and Octasiano Miguel Valerio Mendoza. 2023. "Job prospects and labour mobility in China." *The Journal of International Trade & Economic Development*, 32(7): 991–1034.

Wesolowski, Amy, Caroline O Buckee, Kenth Engø-Monsen, and Charlotte Jessica Eland Metcalf. 2016. "Connecting mobility to infectious diseases: the promise and limits of mobile phone data." *The Journal of infectious diseases*, 214(suppl_4): S414–S420.

Yang, Peiyu, Tongyu Zhu, Xuejin Wan, and Xuejiao Wang. 2014. "Identifying significant places using multi-day call detail records." 360–366, IEEE.

Yuan, Yihong, Martin Raubal, and Yu Liu. 2012. "Correlating mobile phone usage and travel behavior–A case study of Harbin, China." *Computers, Environment and Urban Systems*, 36(2): 118–130.

Zagheni, Emilio, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. "Inferring international and internal migration patterns from twitter data." 439–444.

doi:10.6342/NTU202502893

Zhao, Zhiyuan, Shih-Lung Shaw, Ling Yin, Zhixiang Fang, Xiping Yang, Fan Zhang, and Sheng Wu. 2019. "The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data." *International Journal of Geographical Information Science*, 33(7): 1471–1495.

Zhao, Ziliang, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Ling Yin. 2016. "Understanding the bias of call detail records in human mobility research." *International Journal of Geographical Information Science*, 30(9): 1738–1762.

Zhou, Xiangkai, Linlin You, Shuqi Zhong, and Ming Cai. 2024. "From cell tower location to user location: Understanding the spatial uncertainty of mobile phone network data in human mobility research." *Computers, Environment and Urban Systems*, 111: 102130.

Zreikat, Aymen I, Khalid Al-Begain, and Kevin Smith. 2004. "A comparative capacity/coverage analysis for CDMA cell in different propagation environments." *Wireless Personal Communications*, 28(3): 205–231.

doi:10.6342/NTU202502893

Appendix A — Event-Centered Trends Across Outcomes and Treatments

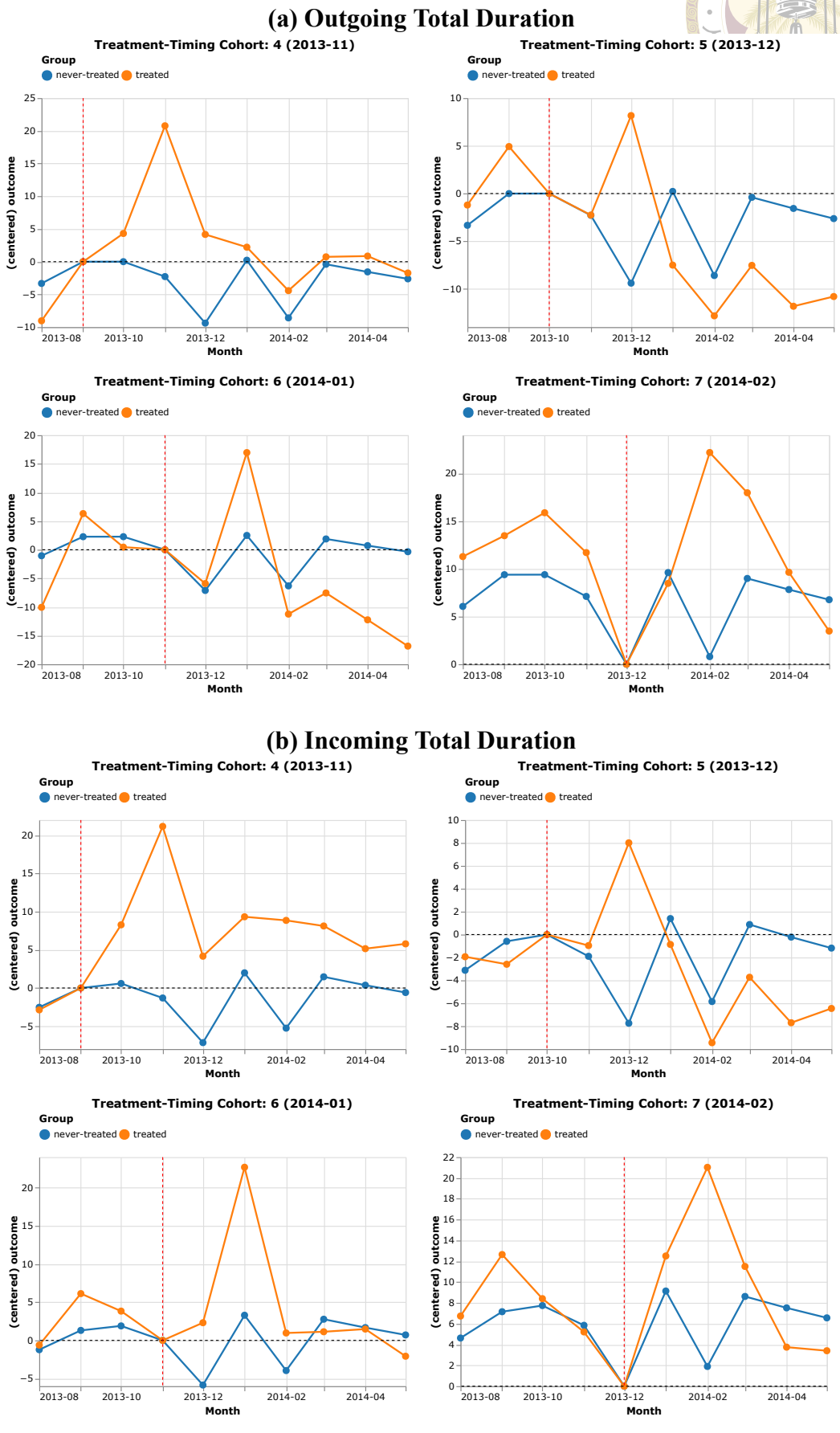
In this section of appendix, we attach all outcomes' mean trajectories over time across two groups of features: mobility and mobile communication network features, and two types of treatments: residential shift and smartphone adoption. These plots should serve as the motivation to discuss the treatment effects as we can clearly observe the substantial and sudden changes after the exposure to treatments across treated groups. Besides, they also validate the plausibility of applying the DiD framework as through examining the outcome trends, where the comparison is made to the reference period, $g - \delta - 1$, we can see nearly identical unconditional trends between the treated and control groups most of the time. The trends we refer heavily here are computed through the re-centering scheme on each sample average outcome $\hat{\mathbb{E}}[Y_{i,m} \mid G_{i,g} = 1]$ for cohort $\mathcal{G}_g$ given δ:

$$\tilde{Y}_m(g) = \hat{\mathbb{E}}[Y_{i,m} \mid G_{i,g} = 1] - \hat{\mathbb{E}}[Y_{i,g-\delta-1} \mid G_{i,g} = 1]$$

where $g \in \{4, 5, 6, 7\}$ or is equal to ∞ while calculating the control group's outcome trends. Therefore, by comparing $\tilde{Y}_m(g)$ to $\tilde{Y}_m(\infty)$ for all $m < g - \delta - 1$, we can assess on how likely parallel trends hold in the pre-treatment periods.

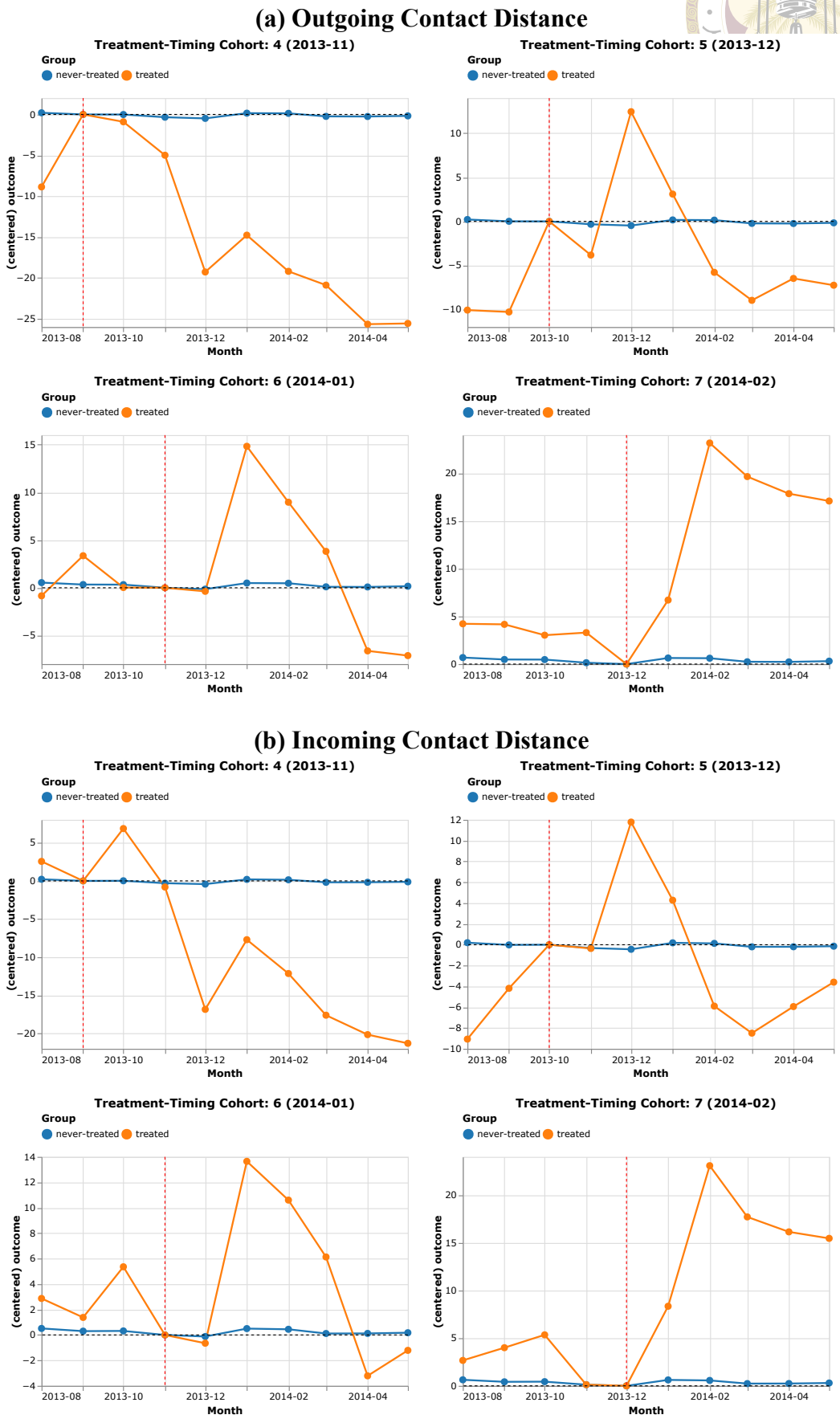
doi:10.6342/NTU202502893

Figure A.1: Total Duration by Treatment (Residential Shifts) Status



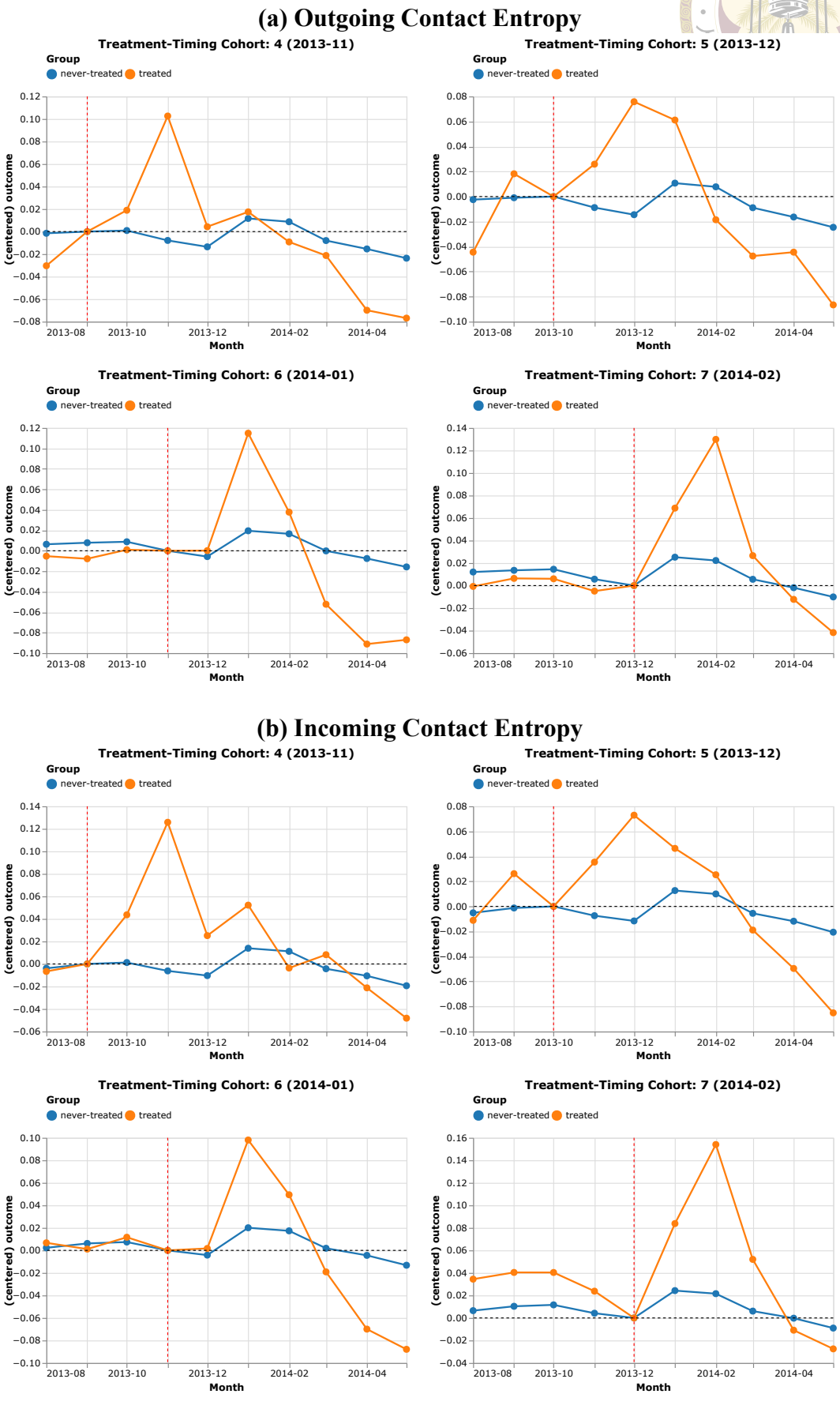
doi:10.6342/NTU202502893

Figure A.2: Contact Distance by Treatment (Residential Shifts) Status



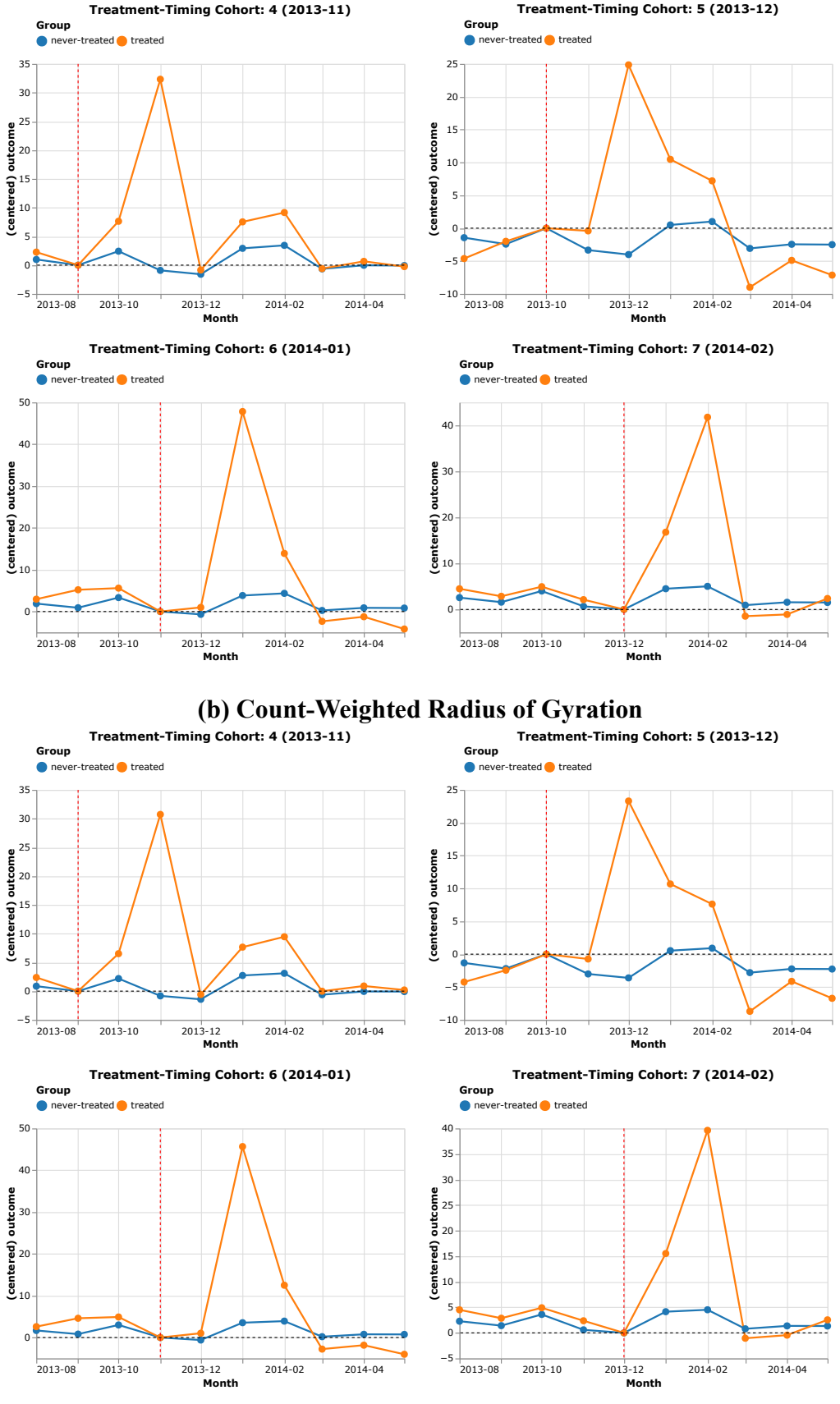
doi:10.6342/NTU202502893

Figure A.3: Contact Entropy by Treatment (Residential Shifts) Status



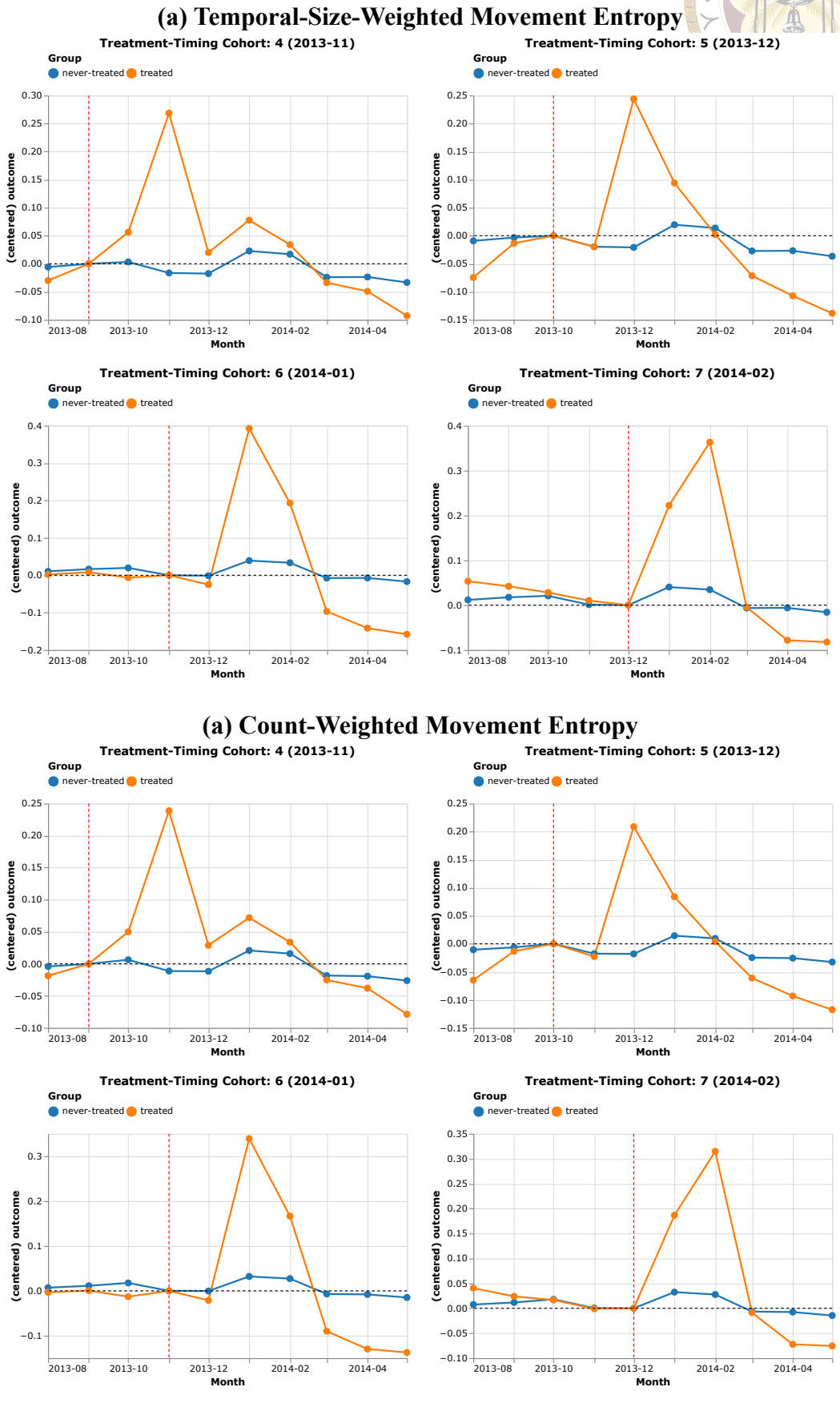
doi:10.6342/NTU202502893

Figure A.4: Radius of Gyration by Treatment (Residential Shifts) Status



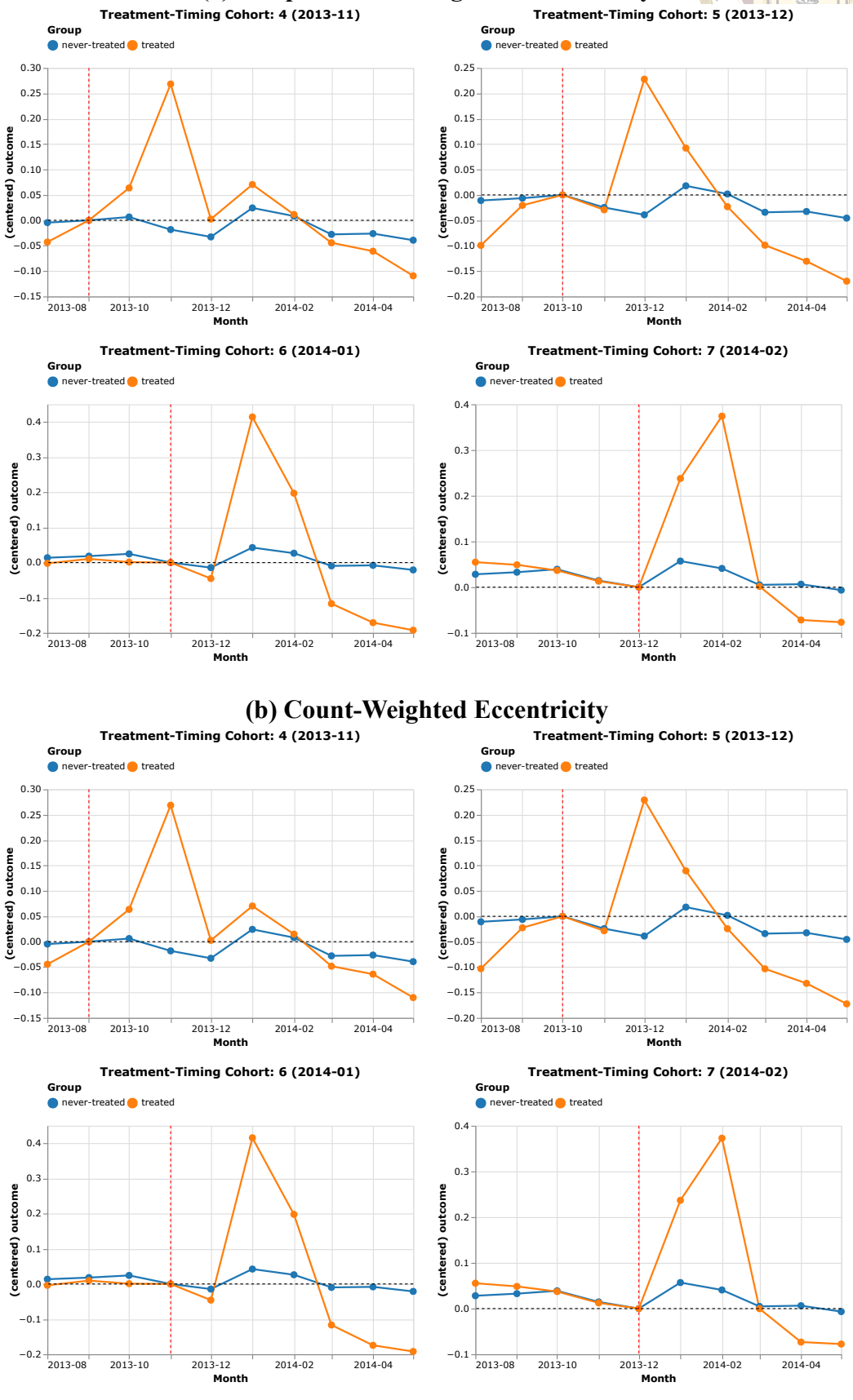
doi:10.6342/NTU202502893

Figure A.5: Movement Entropy by Treatment (Residential Shifts) Status



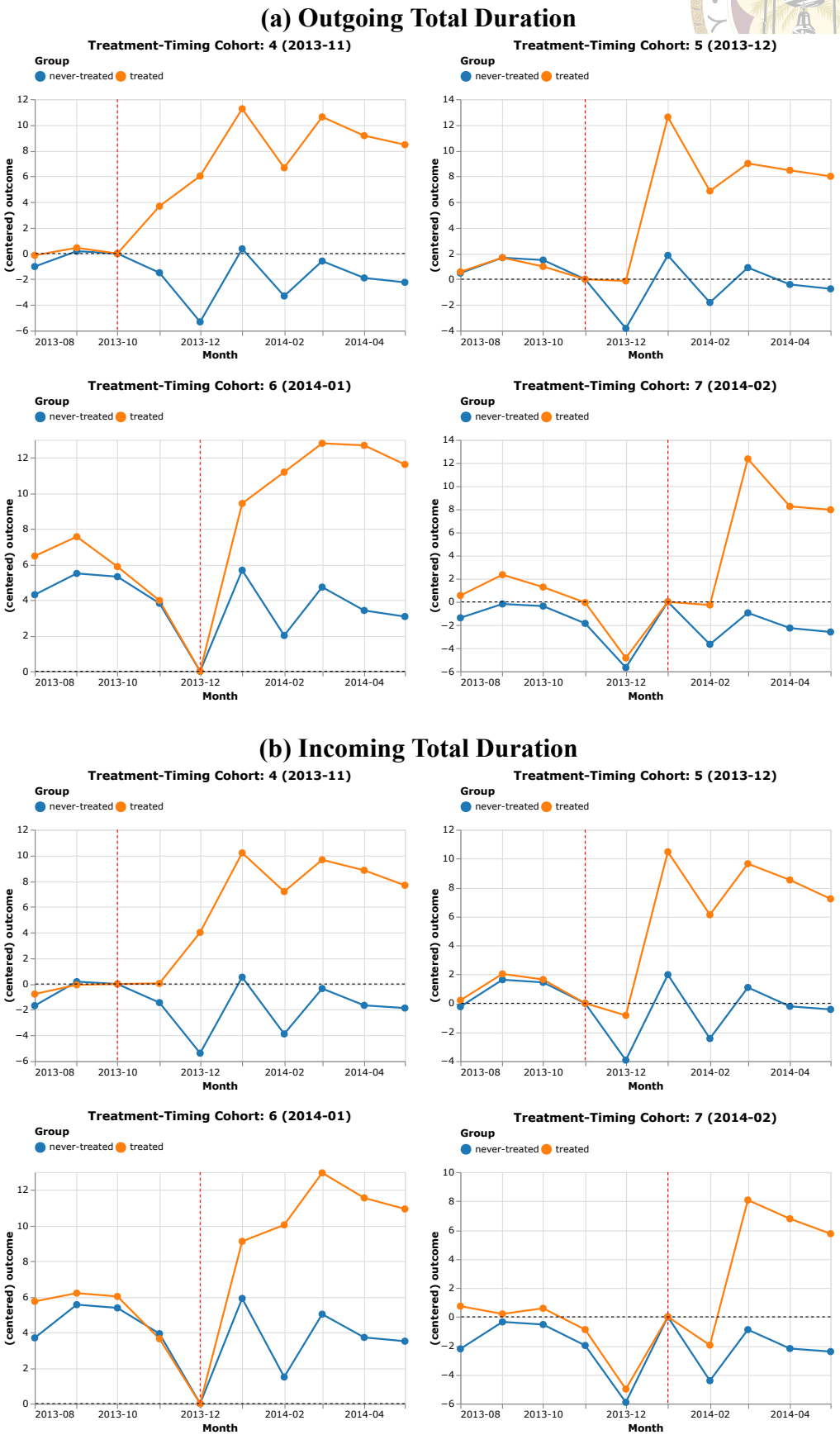
doi:10.6342/NTU202502893

Figure A.6: Eccentricity by Treatment (Residential Shifts) Status



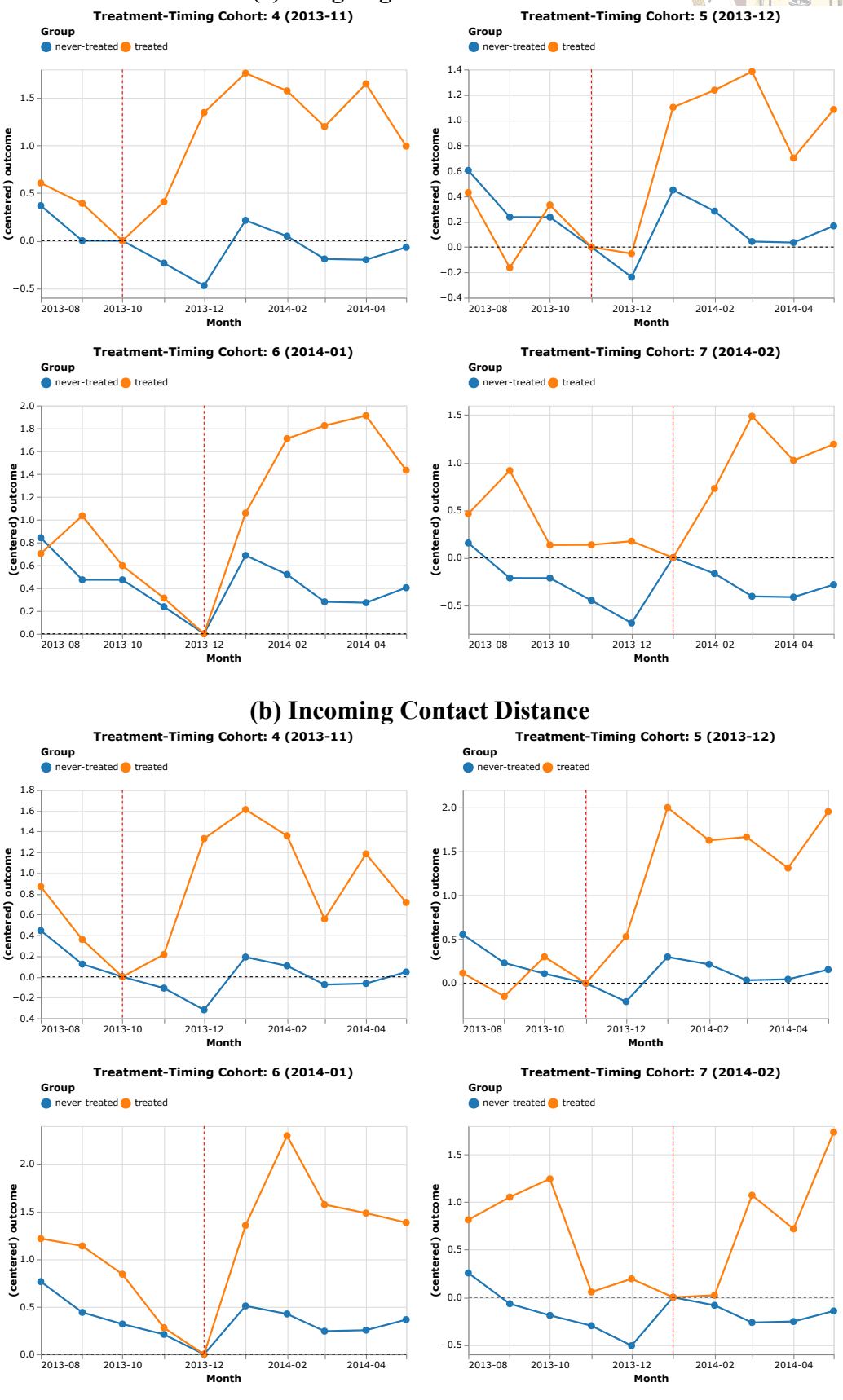
doi:10.6342/NTU202502893

Figure A.7: Total Duration by Treatment (Smartphone Adoption) Status



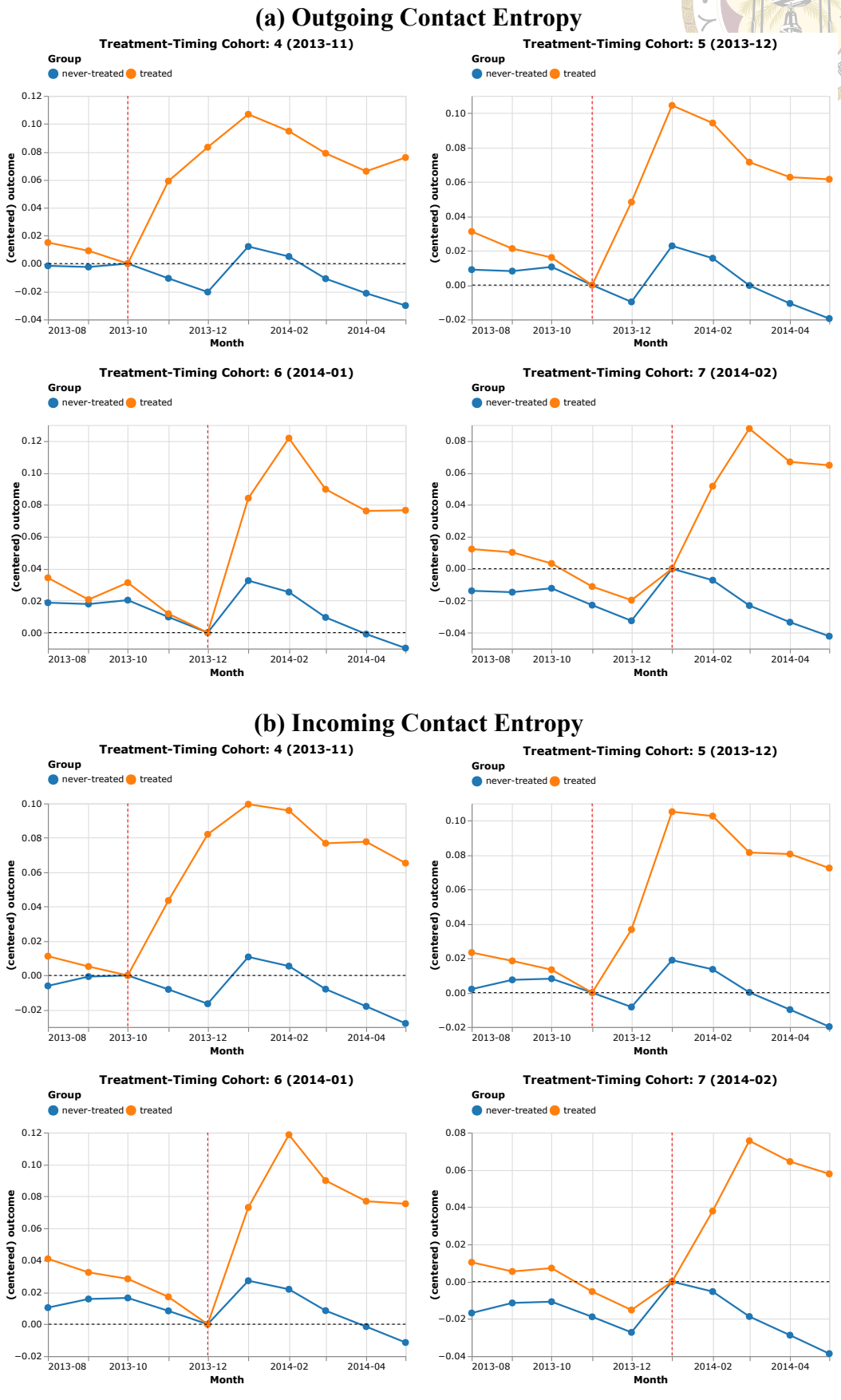
doi:10.6342/NTU202502893

Figure A.8: Contact Distance by Treatment (Smartphone Adoption) Status



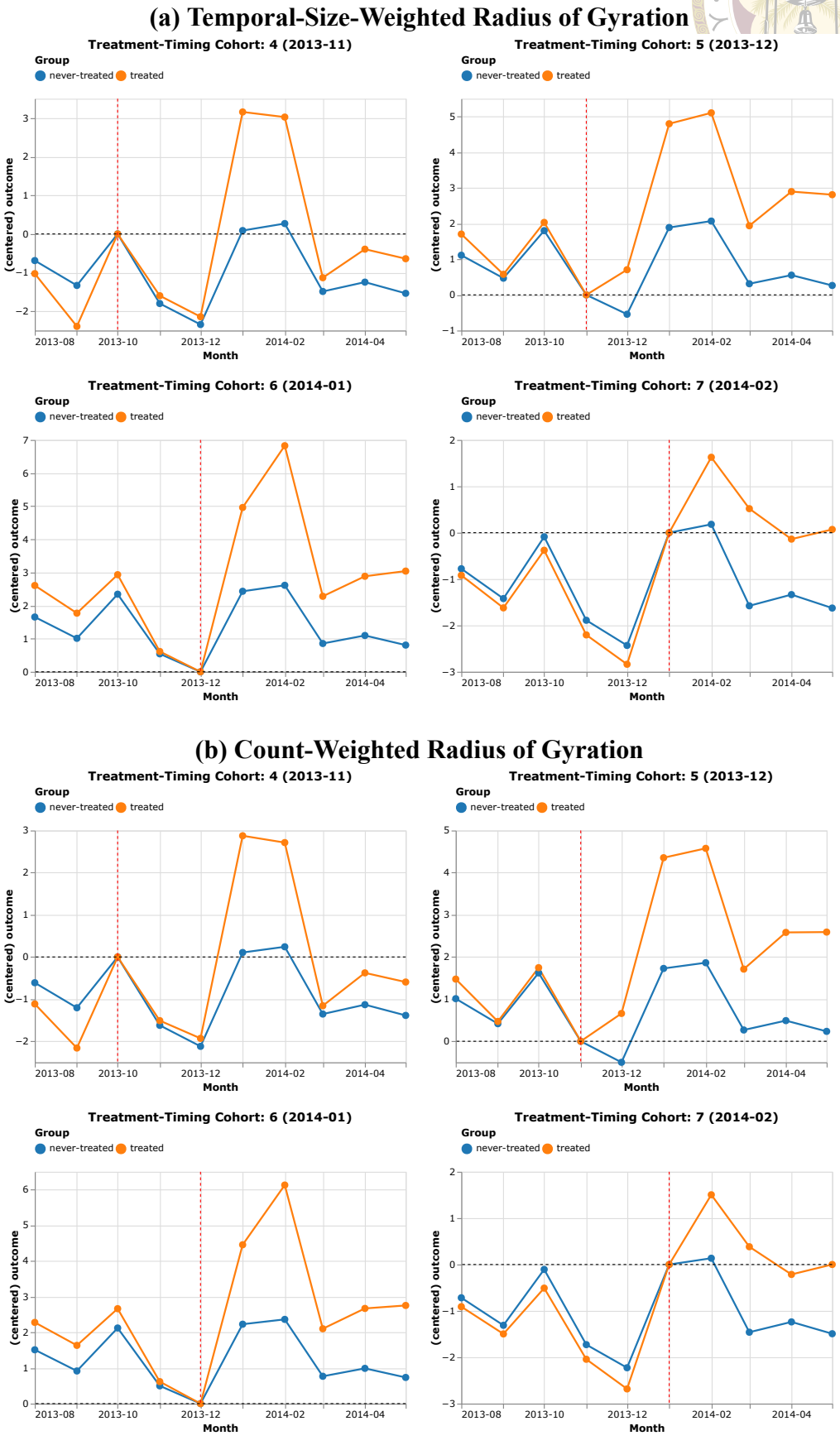
doi:10.6342/NTU202502893

Figure A.9: Contact Entropy by Treatment (Smartphone Adoption) Status



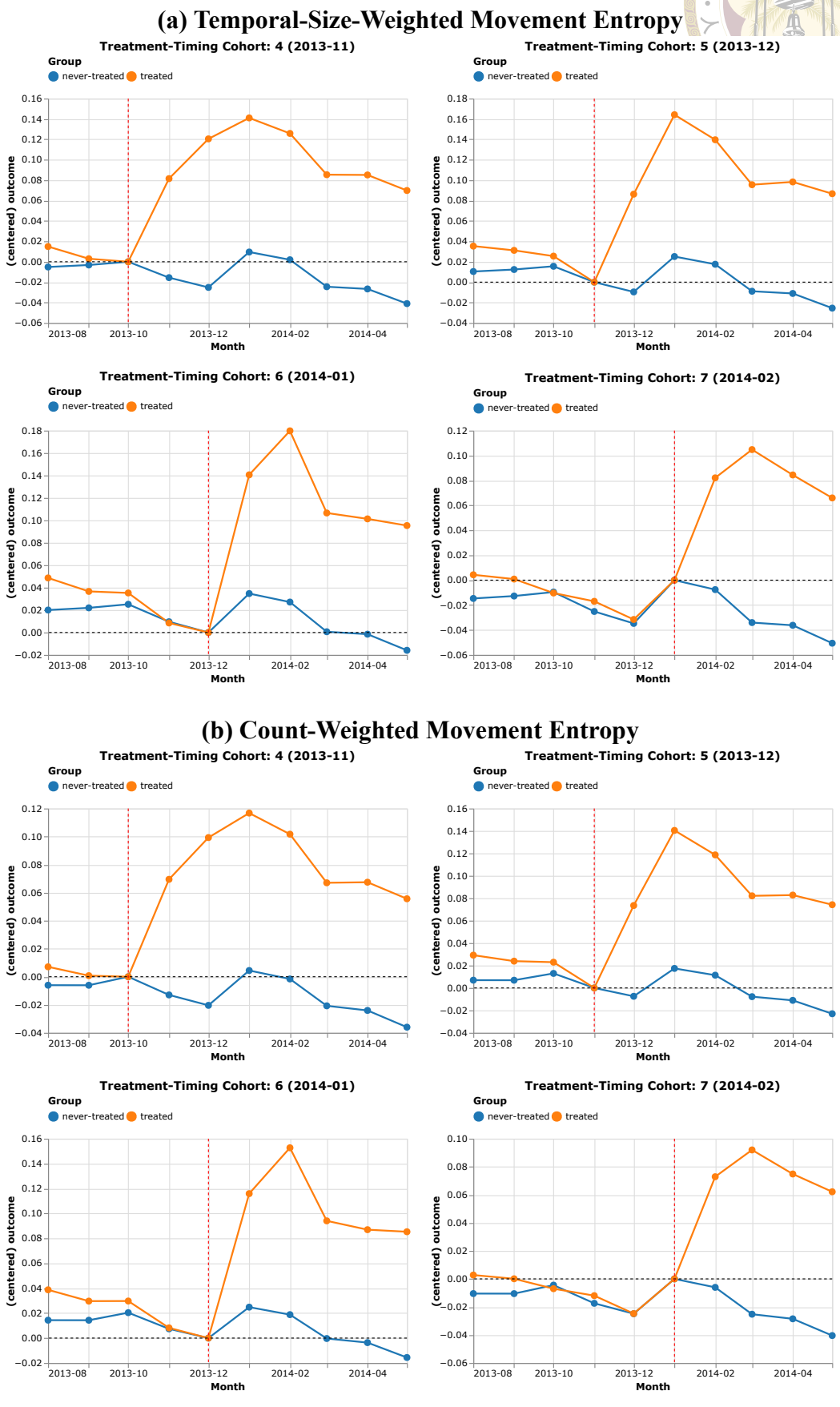
doi:10.6342/NTU202502893

Figure A.10: Radius of Gyration by Treatment (Smartphone Adoption) Status



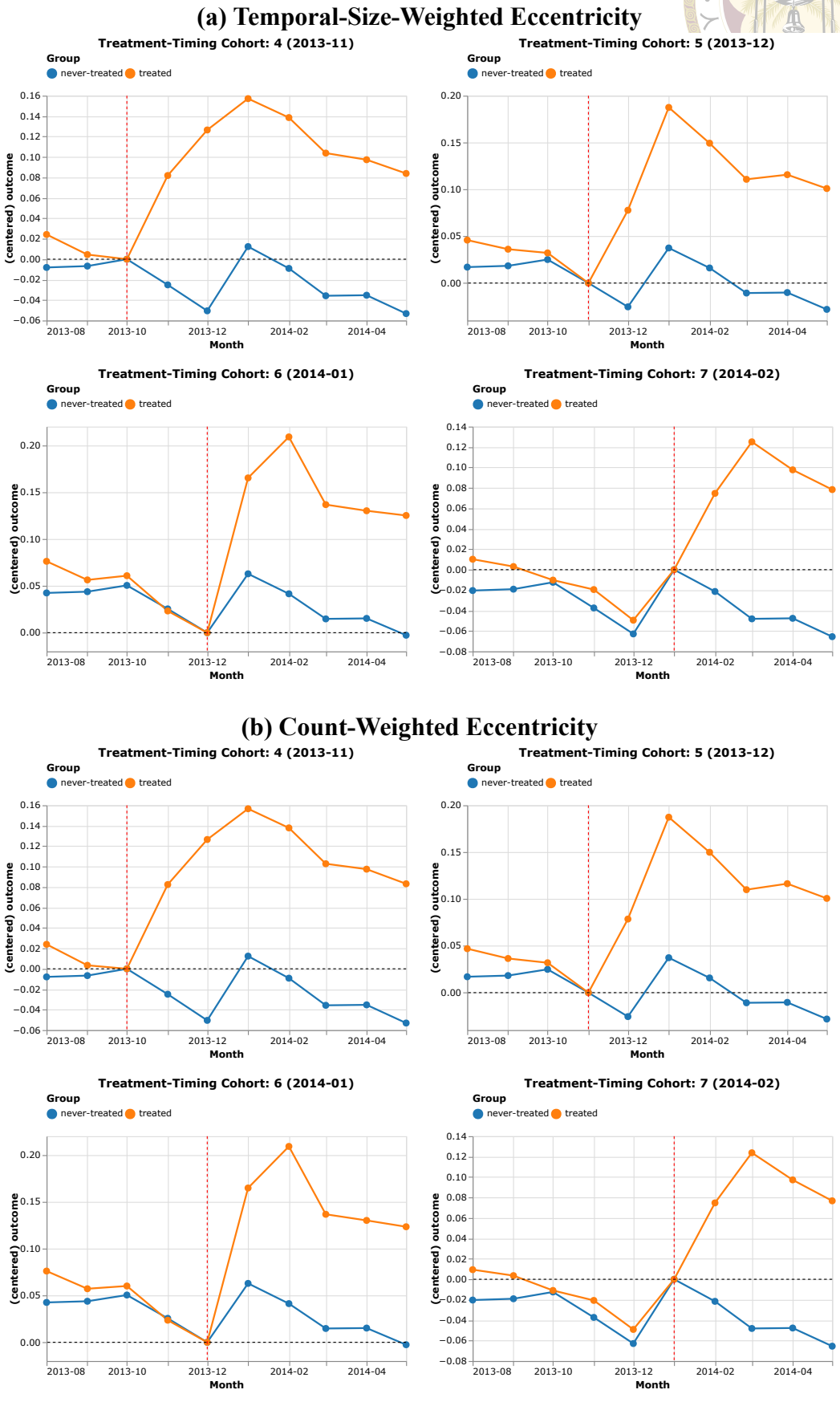
doi:10.6342/NTU202502893

Figure A.11: Movement Entropy by Treatment (Smartphone Adoption) Status



doi:10.6342/NTU202502893

Figure A.12: Eccentricity by Treatment (Smartphone Adoption) Status



doi:10.6342/NTU202502893

Appendix B — Preliminary of DiD Estimator

In this section, we introduce the concept of potential outcome framework and the average treatment effect on the treated (ATT) to formalize our research question. As a straightforward and ideal definition, ATT provides clear guidance on what should be estimated, and it could be obtained through the DiD design. We will elaborate on what specific estimation approach we adopt and the corresponding motivations in this section. Moreover, the estimation approach comes with various possible setups, and we explain which are the most suitable for us.

As we are interested in the effects of residential shifts and smartphone adoption on human mobility patterns and mobile communication behaviors, we formalize our research questions as follows: what are the magnitudes of differences in mobility and communication behaviors for individuals who receive these treatments compared to a counterfactual scenario where they never experienced them? The magnitudes of differences are so-called treatment effects, and treatments in this study are either residential shifts or smartphone adoption events. Given that many individuals in our sample receive the treatment, it is natural to focus on average treatment effects rather than individual-level effects. Moreover, since our primary interest lies in understanding how these life events specifically

doi:10.6342/NTU202502893

impact those who experience them—addressing the counterfactual question "what would have happened if they had not been treated?"—we focus on the treatment effects on the treated specifically as our primary causal parameter of interest.

ATT can be formalized as follows. Suppose a treatment occurs in period t. ATT is given by:

$$ATT = \mathbb{E}[Y_{i,t} - Y_{i,t}(\infty)|D_i = 1]$$

where D_i is a binary variable indicating whether an individual i is treated, $Y_{i,t}$ is the observed outcome in period t, and $Y_{i,t}(\infty)$ is the counterfactual untreated outcome. Examining how ATT evolves over time provides an additional dimension for analysis, and given these motivations, DiD with the design of multiple periods (also known as the event study) is an ideal econometric approach.

The intuition of DiD is that simply comparing treated and untreated units at a single point in time may be misleading because these groups might differ in unobservable ways. Similarly, comparing the same units before and after treatment might confound the treatment effect with general time trends that would have occurred regardless of treatment. DiD solves this problem by using a "double comparison." Suppose the treatment occurs in time period t, and first, it compares the change in outcomes for the treated group over time:

$$\mathbb{E}[Y_{i,t} - Y_{i,t-1} \mid D_i = 1]$$

Second, it compares this change to the change observed in a control group over the same period:

$$\mathbb{E}[Y_{i,t} - Y_{i,t-1} \mid D_i = 1] - \mathbb{E}[Y_{i,t} - Y_{i,t-1} \mid D_i = 0]$$

Besides, what makes the DiD design prevalent in empirical study is that it has good theoret-

doi:10.6342/NTU202502893

ical implication in that if we assume parallel trends in counterfactual untreated outcomes:

$$\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) \mid D_i = 1] - \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) \mid D_i = 0],$$

then the double comparison can recover ATT:

$$\begin{aligned}
&\underbrace{\mathbb{E}[Y_{i,t} - Y_{i,t}(\infty) | D_i = 1]}_{ATT} \\
&= \mathbb{E}[Y_{i,t} \mid D_i = 1] \\
&\quad - \underbrace{(\mathbb{E}[Y_{i,t-1}(\infty) \mid D_i = 1] + \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) \mid D_i = 1])}_{=\mathbb{E}[Y_{i,t}(\infty) | D_i = 1]} \\
&= \mathbb{E}[Y_{i,t} - Y_{i,t-1}(\infty) \mid D_i = 1] - \underbrace{\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) \mid D_i = 0]}_{\text{due to parallel trends}} \\
&= \mathbb{E}[Y_{i,t} - Y_{i,t-1} \mid D_i = 1] - \mathbb{E}[Y_{i,t} - Y_{i,t-1} \mid D_i = 0].
\end{aligned}$$

Note that observed outcomes are equivalent to counterfactual untreated outcomes for all untreated units in both period $t-1$ and t, and the equivalence also holds for treated units at period $t-1$.

doi:10.6342/NTU202502893

Appendix C — Results of ATT Estimation by Event Time

Here we present the full ATT estimation results on which Figure 4.1 and Figure 4.3 are based. We discuss the effects of residential shift and smartphone adoption on mobile communication networks and mobility features. Due to the limited space, we abbreviate part of the variables' names as follows: "out" indicates the outgoing direction of mobile communication from which communication features are derived. "dura" represents the total duration of all phone calls in a month (measured in minutes). "cd" is the abbreviation for contact distance, which is the average geographical distance between phone users and their friends (measured in kilometers). "entr" stands for entropy and there are two types of entropy: out ce (outgoing contact entropy) and ts. me (temporal-size-weighted movement entropy), which quantify the unpredictability of mobile contacts and visited locations, respectively. "ts." is the short form of temporal-size-weighted, indicating the mobility features are computed through temporal-size-based weights rather than count-based ones. "rg" is the abbreviation for radius of gyration (measured in kilometers), which quantifies how large a user's activity area is. "ecc" is the eccentricity, measuring the ratio of the first and second principal components of users' two-dimensional variance-covariance matrix of spatial distribution.

doi:10.6342/NTU202502893

Table C.1: Results of ATT (Residential Shifts) Estimation by Event Time

	out dura	out cd	out ce	ts. rg	ts. me	ts. ecc
ATT(-3)	0.95	−2.77	−0.01	1.45	−0.01	−0.02
	(1.86)	(1.50)	(0.01)	(0.92)	(0.01)	(0.01)
ATT(-2)						
ATT(-1)	0.75	1.42	0.03**	6.40***	0.07***	0.07***
	(1.81)	(1.73)	(0.01)	(1.11)	(0.01)	(0.01)
ATT(0)	19.18***	13.45***	0.10***	36.57***	0.32***	0.32***
	(2.12)	(2.91)	(0.01)	(1.54)	(0.01)	(0.01)
ATT(1)	3.15	6.15	0.03*	3.62**	0.06***	0.06***
	(2.32)	(3.40)	(0.01)	(1.21)	(0.02)	(0.02)
ATT(2)	−2.23	3.09	−0.02	0.37	−0.04**	−0.05**
	(2.34)	(3.32)	(0.01)	(1.05)	(0.02)	(0.02)
ATT(3)	−5.11*	−1.27	−0.04***	−0.22	−0.06***	−0.08***
	(2.47)	(3.08)	(0.01)	(1.00)	(0.01)	(0.02)
Num. users	292739	292739	292739	292739	292739	292739
Num. anticipation	1	1	1	1	1	1

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

doi:10.6342/NTU202502893

Table C.2: Results of ATT (Smartphone Adoption) Estimation by Event Time

	out dura	out cd	out ce	ts. rg	ts. me	ts. ecc
ATT(-3)	0.76*	0.12	0.01***	0.06	0.01***	0.02***
	(0.38)	(0.19)	(0.00)	(0.23)	(0.00)	(0.00)
ATT(-2)	0.16	0.26	0.01**	−0.25	0.00	0.01*
	(0.37)	(0.16)	(0.00)	(0.20)	(0.00)	(0.00)
ATT(-1)	0.00	0.00	0.00	0.00	0.00	0.00
ATT(0)	3.99***	0.56**	0.06***	1.44***	0.10***	0.10***
	(0.38)	(0.18)	(0.00)	(0.23)	(0.00)	(0.00)
ATT(1)	10.96***	1.26***	0.10***	2.51***	0.14***	0.17***
	(0.47)	(0.20)	(0.00)	(0.25)	(0.00)	(0.01)
ATT(2)	9.40***	1.15***	0.09***	2.15***	0.12***	0.14***
	(0.50)	(0.22)	(0.00)	(0.23)	(0.00)	(0.00)
ATT(3)	9.42***	1.29***	0.09***	1.95***	0.11***	0.13***
	(0.52)	(0.25)	(0.00)	(0.24)	(0.00)	(0.01)
Num. users	91446	91446	91446	91446	91446	91446
Num. anticipation	0	0	0	0	0	0

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

doi:10.6342/NTU202502893

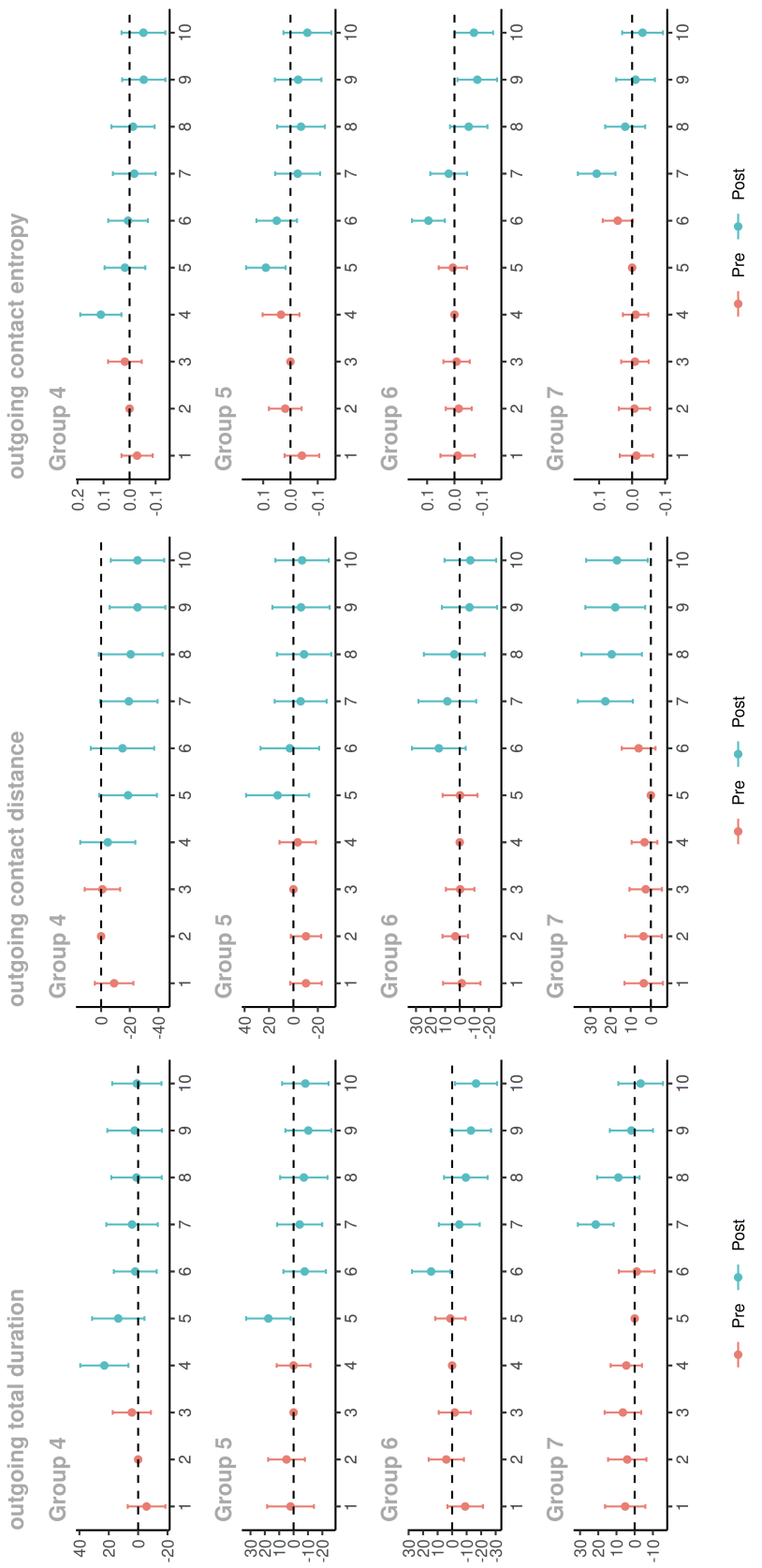
Appendix D — Group-Specific Event Studies

Applying Callaway and Sant' Anna (2021)'s approach involves estimating the ATT for each treatment-timing group and calendar month. Then, a family of causal parameters of interest can be obtained, which can be aggregated through Equation 3.15 to recover the event study. In the main text, we present the event-study-like results, which are based on the aggregated ATT by the number of months after the treatment. One might wonder if there are heterogeneous treatment effects across different treatment-timing groups so we include the complete group-time ATT here for further discussions.

Through the following set of figures, we can generally claim that there is no substantial heterogeneous treatment effects across different treatment-timing groups as there is no particular group exhibit distinct ATT dynamics compared to the others in any of the outcome. However, there is an exception, which is ATT of residential shift on outgoing contact distance in group 7 (see Figure D.13). In such case, only group 7 shows a significant positive effect contemporaneously with the treatment while all the other groups exhibit insignificant effects. However, all the others are, in fact, experience a positive upward shift in outcomes, which coincides with the group 7 and therefore, we can still somehow confirm the positive effect of residential shift on outgoing contact distance.

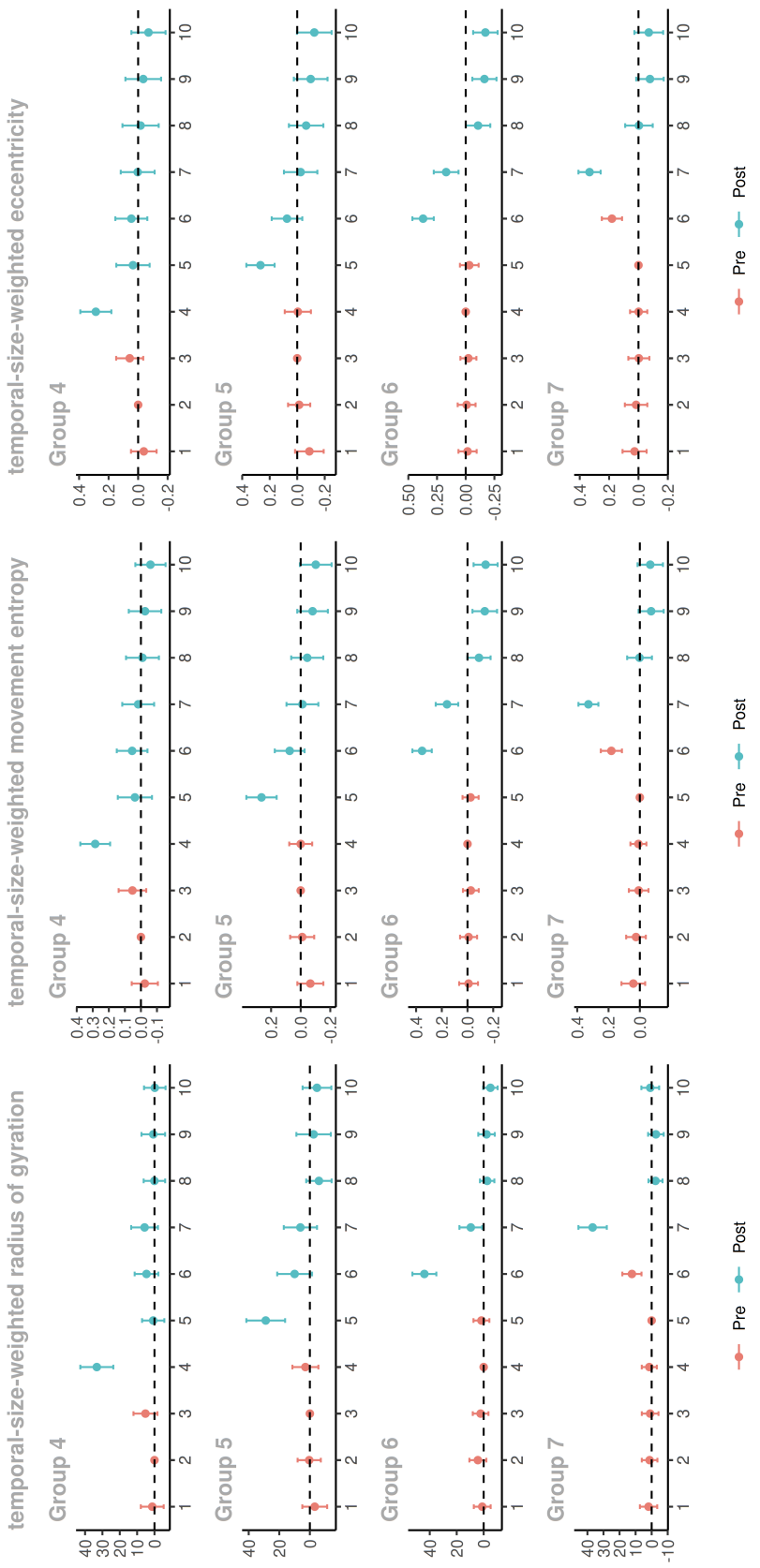
doi:10.6342/NTU202502893

Figure D.13: Group-Specific Event Study: Residential Shifts on Communication



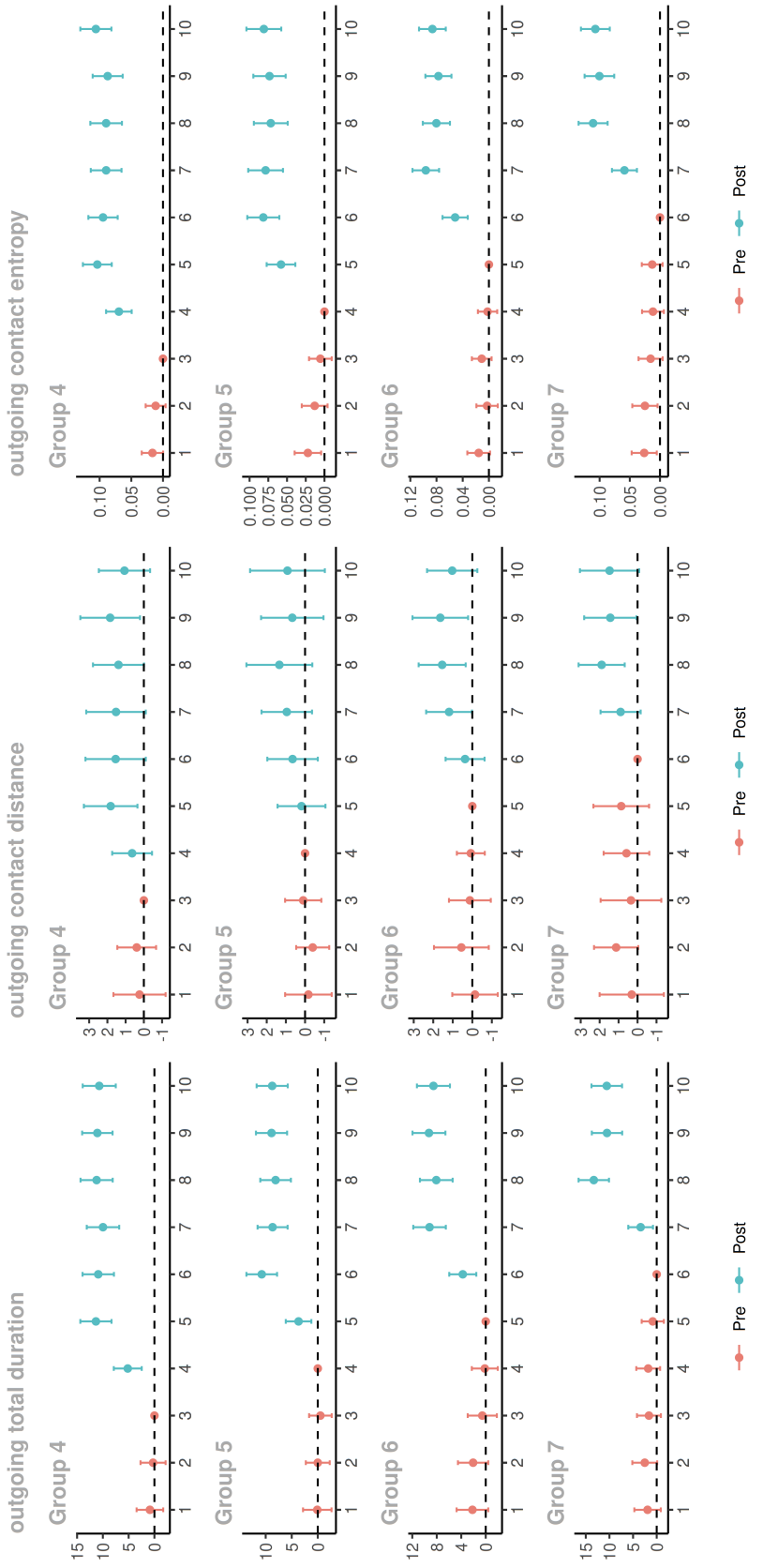
doi:10.6342/NTU202502893

Figure D.14: Group-Specific Event Study: Residential Shifts on Mobility



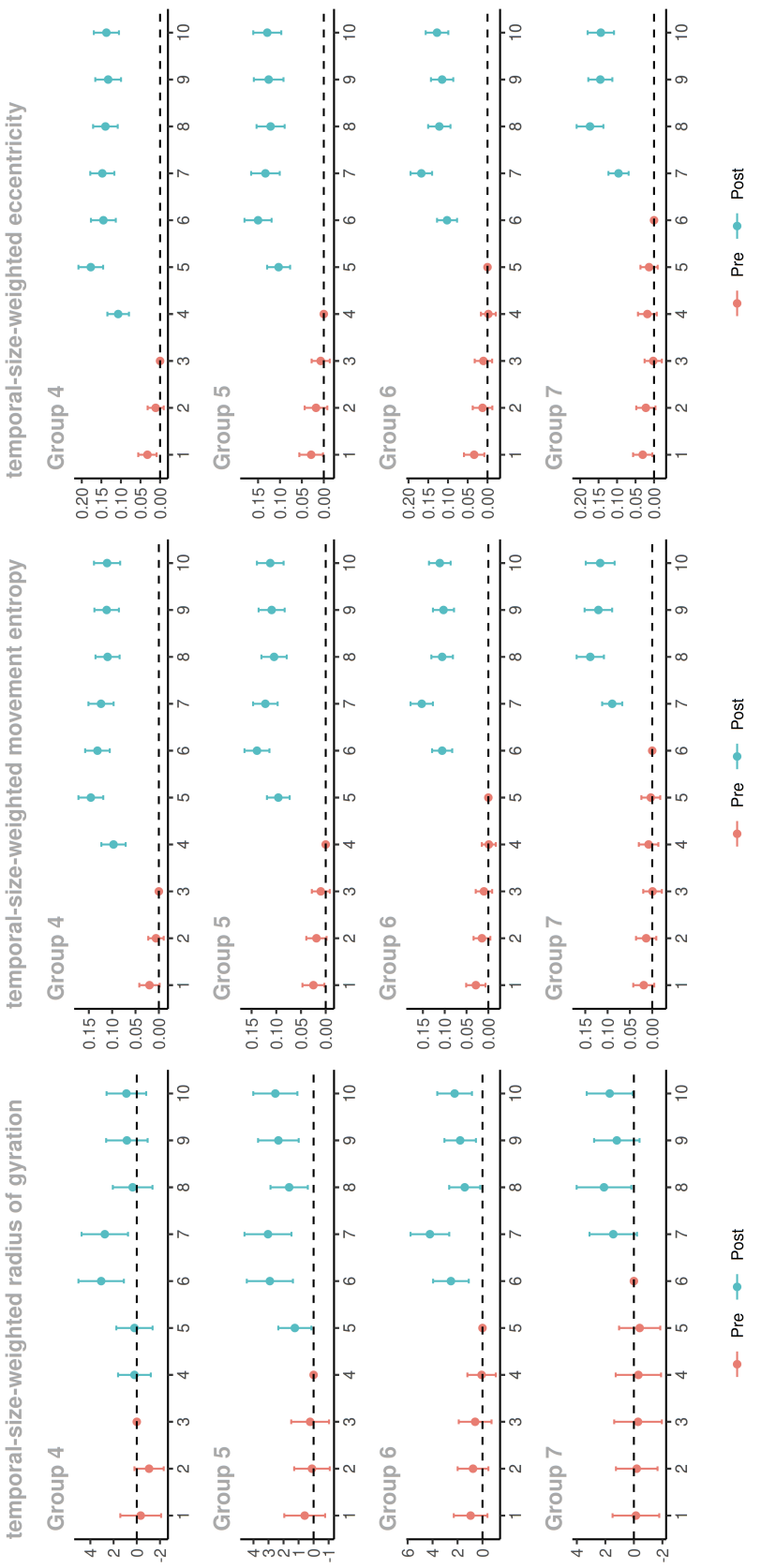
doi:10.6342/NTU202502893

Figure D.15: Group-Specific Event Study: Smartphone Adoption on Communication



doi:10.6342/NTU202502893

Figure D.16: Group-Specific Event Study: Smartphone Adoption on Mobility



doi:10.6342/NTU202502893

Appendix E — Implementation Details

E.1 Parameter Choices of DBSCAN

While most cases involve multiple observed locations forming a cluster around the home location candidate, it is possible that only a single observed location is associated with a home location candidate, particularly in areas with sparse base station coverage. To account for this case, we set the *min_samples* parameter to one. Another parameter, *eps*, which defines the maximum distance for two observed locations to be considered neighbors within a cluster, is set to 5 km with distances calculated using Vincenty's formulae. This parameter choice is motivated by the heterogeneous nature of effective service radii of base stations across our study regions.

Since effective service radii provide informative insights into the neighboring distances between consecutive base stations, they serve as appropriate prior knowledge for determining the *eps* value. Theoretically, neighboring distances should be less than the sum of two consecutive base stations' effective service radii. Therefore, *eps* should be greater than the maximum of all neighboring distances approximated by the sum of consecutive base stations' effective service radii but shouldn't be excessively large, as an overly large value might cause the algorithm to incorrectly merge two distinct clusters into one. Nevertheless, we believe that individuals will stay at home most of the time, so

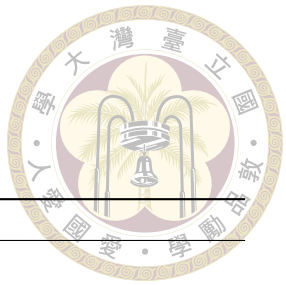
doi:10.6342/NTU202502893

the distance between clusters should be relatively large to mitigate the issue of misidentification of clusters.

Our study regions including Deyang, Chongqing, and other prefectures in Sichuan Province encompass diverse geographic regions, including urban, suburban, and rural areas, and typically, the service radii in urban areas is smaller than those in rural areas (Zreikat, Al-Begain and Smith (2004)). Zhou et al. (2024) provides an overview of recent research that utilizes CDRs to locate individuals' positions across various regions, including a particular discussion on spatial resolution, which is partially related to base stations' service radii. The overview states that the average service radius in urban regions (Shanghai, Nanjing, and Guiyang) is less than 1 km, while our study regions have much more complicated compositions. Therefore, the 5 km threshold represents an aggressive lower bound, which aims to account for the larger service radii characteristic of rural and suburban regions while maintaining meaningful spatial clustering in dense urban regions. Besides, choosing a relatively large *eps* value addresses a key trade-off: while smaller values would reduce localization accuracy in rural regions, larger values risk merging distinct urban clusters. However, this risk can be potentially mitigated because individuals spend most time at home locations, and our weighted-average estimation across telecom stations (based on usage frequency) ensures that wrongly-included stations receive low weights in the final home estimation.

doi:10.6342/NTU202502893

E.2 Temporal Filtering



Algorithm 1 Home Cluster Estimation

Input: C_i^{night}
Output: C_i^{home}

candidates $\leftarrow$ sort C_i^{night} in descending order by the temporal size
$C_i^{\text{home}} \leftarrow$ [candidates[1]]

for all $j = 2, \cdots,$ length(candidates) **do**
  isolate $\leftarrow$ true
  candidate $\leftarrow$ candidates$[j]$

  for all $k = 1, \cdots,$ length(C_i^{home}) **do**
    if candidate temporally overlaps with $C_i^{\text{home}}[k]$ **then**
      isolate $\leftarrow$ false
      break
    end if
  end for

  if isolate $=$ true and the temporal size of candidate > 2 **then**
    insert candidate into C_i^{home}
  end if
end for

sort C_i^{home} in ascending order by the start time of service time interval
return C_i^{home}

Theorem E.2.1. For the algorithm 1, where each user $i \in V$ has a nighttime cluster set C_i^{night} with $|C_i^{\text{night}}|$ elements, the time complexity is $O\left(\sum_{i \in V} |C_i^{\text{night}}|^2\right)$ and the space complexity is $O\left(\sum_{i \in V} |C_i^{\text{night}}|\right)$.

Proof. For the time complexity, consider the worst case scenario where for all user i, all nighttime clusters are temporally non-overlapping with one another. The algorithm will iterate for

$$1 + 2 + \ldots + (|C_i^{\text{night}}| - 1) = \frac{|C_i^{\text{night}}|(|C_i^{\text{night}}| - 1)}{2} = O(|C_i^{\text{night}}|^2)$$

times for each user i. Aggregating over all users gives the total time complexity of

doi:10.6342/NTU202502893

$O\left(\sum_{i\in V}|C_i^{\text{night}}|^2\right)$. For the space complexity, under the worst case scenario, the size of C_i^{home} is identical to $|C_i^{\text{night}}|$ for each user i. Therefore, the total space required is $O\left(\sum_{i\in V}|C_i^{\text{night}}|\right)$. □

E.3 Residential Shifts

Table E.3: Statistics of Migrants by the Month of Migration

	Count of Migrants			Migration Distance		
month	all	inter-pref.	ratio (%)	all	inter-pref.	ratio (%)
Aug. 2013	2434	2006	82.42	134.8	157.11	116.55
Sep. 2013	974	797	81.83	135.82	159.25	117.25
Oct. 2013	451	311	68.96	119.01	158.83	133.46
Nov. 2013	363	247	68.04	101.0	137.35	135.99
Dec. 2013	305	231	75.74	132.15	164.7	124.63
Jan. 2014	448	338	75.45	128.04	160.78	125.57
Feb. 2014	594	458	77.1	116.56	142.44	122.2
Mar. 2014	594	443	74.58	118.48	148.91	125.68
Apr. 2014	675	519	76.89	129.55	160.28	123.72
May 2014	1520	1222	80.39	127.37	151.29	118.78

Notes: The column all represents the count of migrants and migration distance are computed on the phone users who satisfy the first requirement of migrants. Therefore, they don't necessarily change their home locations to another prefecture. The column, inter-pref., means statistics are computed on the phone users who satisfy the first and second requirement of migrants. Moreover, the unit of distance is in kilometers.

The treatment timing varies across users who have once changed their residential locations, but we select those who migrate in the middle of the sample period, i.e., $g \in \{4, 5, 6, 7\}$. For these users, we have higher confidence level to safely classify them as migrants in that the temporal sizes of the two home clusters are comparable. For example, if the residential shift takes place in September 2013, then the first cluster's temporal size is about a month while the second cluster's temporal size is very likely to be greater than a month with a maximum of 9 months. In this case, the two home clusters are obviously not

doi:10.6342/NTU202502893

comparable in terms of the temporal sizes, so we are less confident to classify this user as a migrant because the first cluster may be a short visit, and the first timestamp associated with the second home cluster might be even earlier, outside the sample period. We call this kind of issue observation window bias, which results from the pre-determined observation time frame where the accumulation of information for inference is insufficient. Referring to Table E.3, it seems that there are many users relocating in the early and late period of the sample period, highlighting the importance to restrict the definition of migrants to those relocate in the middle of the sample period. Besides, restricting migrants to those who relocate in the middle of the sample period doesn't substantially distort the origin-destination distribution, as the KL-divergence is about 0.35, which is calculated by comparing the origin-destination distribution of inter-prefecture migrants who migrate in the middle of the sample period to that of migrants who migrate in all months.

We can define some parameters to rule out users who have incomparable temporal sizes of home clusters, for example, requiring the temporal sizes of the home clusters or even the ratio between them to be greater than thresholds. Nonetheless, as aforementioned, it's not necessary to define such parameters. We can simply follow the patterns of the data and conduct restrictive sample selection to validate the robustness of the results. Moreover, even if these parameters are defined, they are not employed to reduce methodological error—like other existing methods to identify residential shift mentioned in Section 2.1. Rather, it's a decision on how much we should trust the patterns

doi:10.6342/NTU202502893

Table E.4: Balance of Pre-Treatment (Residential Shifts) Covariates

Variable	Non-Migrant	Migrant	Difference	p-value
Panel A: Inter-Prefecture Migrants, Any				
age	40.87	36.67	-4.2	0.0
male flag	0.66	0.68	0.02	0.0001
born in Deyang flag	0.82	0.65	-0.17	0.0
(max) phone price	954.72	1059.87	105.15	0.0
(max) smartphone flag	0.65	0.77	0.12	0.0
Panel B: Inter-Prefecture Migrants, Middle Period				
age	40.87	35.56	-5.31	0.0
male flag	0.66	0.71	0.05	0.0001
born in Deyang flag	0.82	0.61	-0.21	0.0
(max) phone price	954.72	1011.0	56.28	0.0234
(max) smartphone flag	0.65	0.74	0.09	0.0

Notes: (i) Panel A presents statistics for phone users who meet the first two requirements of the migrant definition outlined in Definition 3.4.1 without restricting the migration timeframe, while Panel B analyzes data based on the complete migrant definition, limiting inter-prefecture migrants to those who migrated between November 2013 and February 2014. (ii) The variables include: male flag, a binary indicator of user gender; born in Deyang flag, a binary variable indicating whether the user was born in Deyang prefecture; and (max) phone price, and (max) smartphone flag, which are time-variant variables constructed using data before November 2013 to examine pre-treatment covariates. (iii) Since phone users may have changed devices or own multiple phone devices between August 2013 and October 2013,(max) phone price represents the highest price among all phones a user owned during this period, and (max) smartphone flag indicates whether a user ever owned a smartphone during this period.

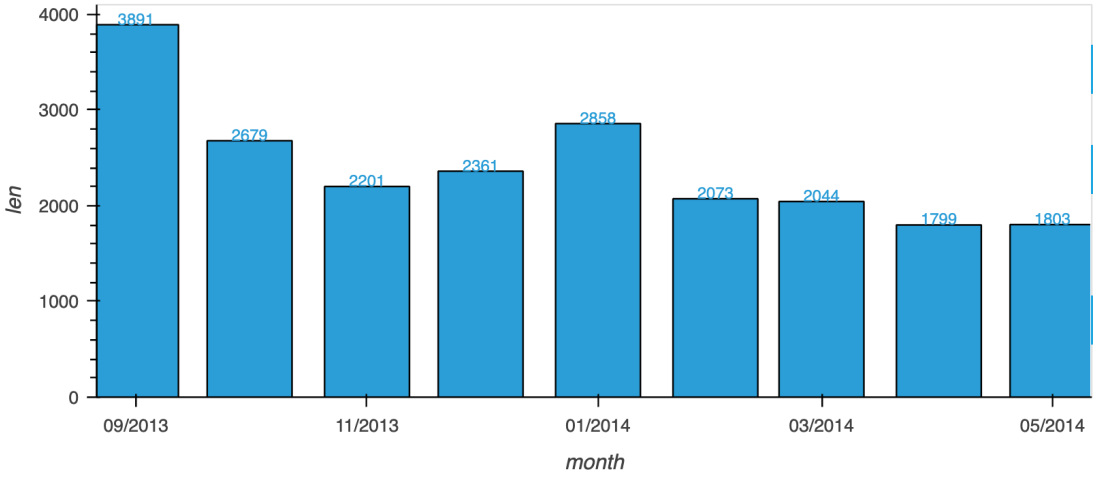
Table E.4 presents the sample statistics of our final selection on migrants, compared to the non-migrant groups. We can see that migrants' characteristics differ slightly between the complete migrant sample and the subsample including only those who migrate in the middle of sample period. The differences in demographic features compared to non-migrants are larger for the subsample migrants than for the complete migrant sample, while the differences in phone-related characteristics compared to non-migrants are smaller for the subsample migrants.

doi:10.6342/NTU202502893

Examining Panel B more closely, as we define the treatment group for analysis as those who migrate across prefectures during the middle of the sample period. Compared to non-migrants, these migrants tend to be younger and have a higher probability of being male, with a lower likelihood of being born in Deyang city. Furthermore, they own slightly better phone devices and a higher fraction of them use smartphones. Although the imbalance of pre-treatment covariates is significant, the scale of differences seems to be small.

E.4 Smartphone Adoption

Figure E.17: Number of Phone Users Upgrading to Smartphones by Month



The above figure shows the counts of smartphone adopters for each month. Generally, there are few differences across months, except a notable increase in September 2013, owing to the observation window bias.

Table E.5 shows that smartphone adopters who upgrade during middle periods are largely similar to those upgrading during any month of the sample period, with the exception of owning slightly less expensive devices. This pattern may result from observation window bias. Panel A includes users who switch to smartphones during early sample

doi:10.6342/NTU202502893

periods, but we lack sufficient observation periods to verify their consistent use of non-smartphone devices before adoption.

The definition of "changers" in Panel B of Table E.5 corresponds to our formal definition of smartphone adopters presented in Definition 3.5.1. We can see that smartphone adopters are slightly younger, marginally more likely to be born outside Deyang prefecture (by approximately 2%), and own more expensive non-smartphone phone devices (before adoption) compared to non-smartphone users. Besides, the age composition of the two groups shows no significant difference. Similar to the situation in residential shifts, the imbalance seems to be not obvious.

Table E.5: Balance of Pre-Treatment (Smartphone Adoption) Covariates

Variable	Non-Changers	Changers	Difference	p-value
Panel A: Smartphone Adopters, Any				
age	44.7	41.68	-3.02	0.0
male flag	0.66	0.65	0.0	0.5039
born in Deyang flag	0.86	0.83	-0.03	0.0
(max) phone price	409.9	706.11	296.22	0.0
Panel B: Smartphone Adopters, Middle Periods				
age	44.7	41.84	-2.86	0.0
male flag	0.66	0.66	0.0	0.8563
born in Deyang flag	0.86	0.83	-0.02	0.0
(max) phone price	409.9	501.75	91.85	0.0

Notes: (i) The Changers refers to the smartphone adopters and non-changers mean non-smartphone users. (ii) The Panel A compares pre-treatment covariates between smartphone adopters (upgrading in any month) and non-smartphone users. Panel B examines the same comparison but restricts smartphone adopters to those upgrading between November 2013 and February 2014. (iii) The pre-treatment covariates are crafted in the same way with Table E.4.

doi:10.6342/NTU202502893

E.5 Selection of Anticipation Parameter

In this section, we explain how we determine the anticipation parameter δ, which is the number of months allowed for anticipation. As we aim to identify effects of residential shift and smartphone adoption on two groups of outcomes—mobility and mobile communication network features—it is plausible for migrants to change their mobility and mobile communication behavior prior to their relocations, as explained previously. However, it's subtle whether users will anticipate upgrading their devices to smartphones. Therefore, we will primarily focus on the residential shift as an illustration example and apply the strategy developed in this section on both treatment contexts.

To select the correct horizon of δ, we initiate a warm-up estimation for the group-time ATT by applying the Callaway and Sant' Anna (2021)'s method implemented in the *did* R package.[1] Several critical settings include setting the *anticipation* argument to 0, corresponding to $\delta = 0$ and the *base_period* argument to "varying". Moreover, we rely on the conditional parallel trend assumption and the never-treated control units. By setting the *anticipation* argument to 0, the post-treatment estimation on group-time ATT is referred to the one period (month) prior to residential shift, and the "varying" *base_period* allows the group-time ATT to be estimated in the reference period, unlike the conventional event study design[2].

In traditional event study, outcomes in both post-treatment and pre-treatment periods are compared to the reference period, which is the one period prior to the treatment if there is no anticipation. Therefore, the reference period cannot compare to itself, resulting in

[1] The *did* R package, to which both authors of the paper, Callaway and Sant' Anna (2021), have been contributing.

[2] The *did* R package allows the event study design by setting *base_period* to "universal".

doi:10.6342/NTU202502893

the failure of estimation. Nonetheless, "varying" *base_period* means that for each period before the treatment, the group-time ATT is estimated through iteratively changing the reference period. That is, to estimate the ATT in one month prior to the treatment, which is usually the reference period, the two periods prior is employed as reference.

"Varying" *base_period* is plausible in pre-treatment periods as it relies on the PTA specific to two-period DiD (see Equation 3.11), which involves the short difference, i.e., $Y_{i,m} - Y_{i,m-1}$, instead of the long difference $Y_{i,m} - Y_{i,g-\delta-1}$ utilized in the post-treatment estimation of DiD with multiple periods (see Equation 3.12) within each treatment group's estimation of ATT dynamics. The motivation for the long difference is that $Y_{i,m-1}(\infty)$ is still unobservable in post-treatment periods if $m - 1 \neq n - \delta - 1$ whereas this does not hold in the pre-treatment periods. Therefore, the short difference is sufficed and applied.

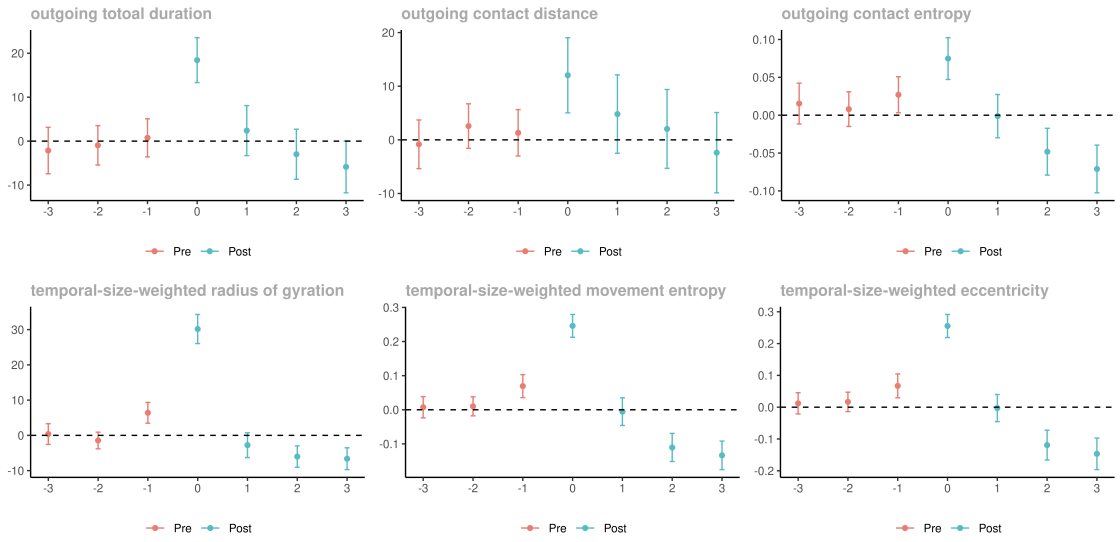
The reason why we specifically want the group-time ATT to be estimable in the reference period $g-\delta-1$ is that it is the most likely period for treatment cohorts to anticipate. Furthermore, iteratively changing the reference period allows us to more easily observe the jump in the plot of group-time ATT dynamics during the pre-treatment period, thereby hypothesizing the occurrence of the anticipation behavior. Moreover, It's relatively computationally efficient compared to conventional event study in that we only need one estimation procedure by setting *base_period* to "varying" to have complete ATT estimation in all pre-treatment periods. Nevertheless, since event study can't estimate ATT in the reference period, we may need to try out different anticipation values, running several estimation procedures to find the right one.

One can view the estimation of ATT in the pre-treatment periods as the placebo test, which aims to answer a hypothetical question: what is the treatment effects if the users

doi:10.6342/NTU202502893

pretendedly receive the treatment prior to the real treatment timing? If the treatment is not confounded, anticipation parameter δ is correctly specified, and PTA holds, the placebo test should yield an insignificant ATT.

Figure E.18: Aggregate Event Study of Residential Shifts with No Anticipation



In Figure E.18, we plot the warmup estimation results of group-time ATT of residential shift on two groups of features. We can see that the group-time ATT is very often significantly different from 0 in the one-month prior to the residential shift and the ATT in period $g-1$ (event time -1) is in the same direction with the period g (event time 0). Note that we employ the 95% confidence bands. Therefore, we think that the number of months for anticipation δ should be set to one when discussing treatment effects of residential shift.

Note that for some outcomes, such as outgoing duration and contact distance, it doesn't make sense to claim the existence of anticipation as ATT in event time -1 (period $g-1$) denoted as $ATT(-1)$ is insignificant. However, it won't affect the estimation of post-treatment effect on these outcomes when requiring the anticipation months to be one, i.e., $g-2$ is referenced. This is because as the "varying" *base_period* let $ATT(-1)$

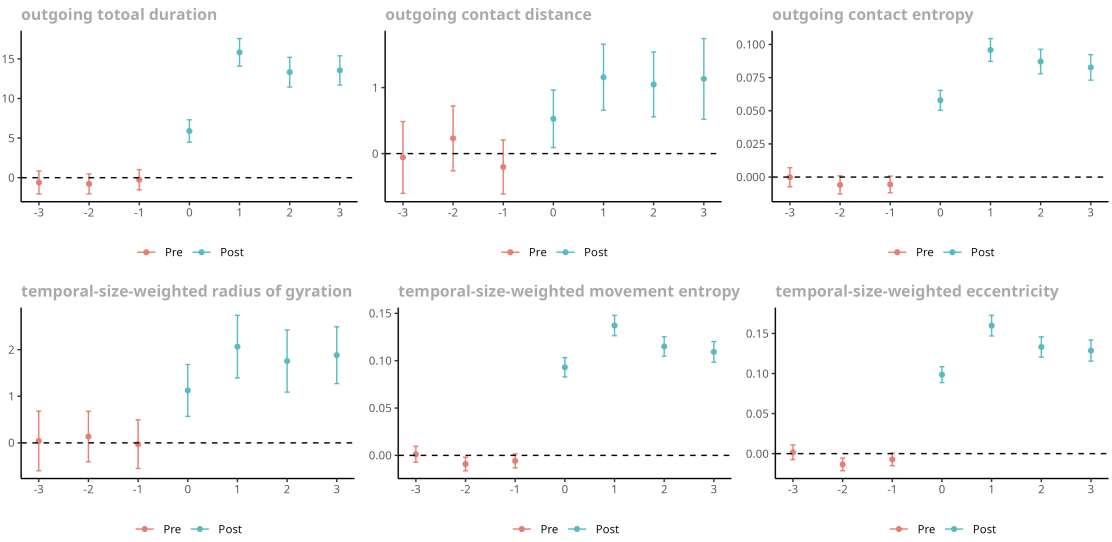
doi:10.6342/NTU202502893

be obtained by referencing period $g-2$, and as the plot shows, $ATT(-1)$ is insignificant, which means the difference, $\mathbb{E}[Y_{i,g-1} - Y_{i,g-2} \mid G_{i,g} = 1]$ is equivalent to the parallel trend. Therefore, it will be differenced out by $\mathbb{E}[Y_{i,g-1} - Y_{i,g-2} \mid G_{i,\infty} = 1]$.

Nevertheless, the situation is not symmetric for outcomes other than outgoing duration and contact distance when we incorrectly set $\delta = 0$ while the true value is $\delta = 1$. This asymmetry arises because $Y_{i,g-1}$ contains an ATT component, and differencing other periods' outcomes against this contaminated baseline distorts the ATT estimates in those periods. Specifically, the ATT in other periods will be underestimated when treatment effects across periods have the same sign, but amplified when they have opposite signs.

Regarding the treatment of smartphone adoption, it seems to be unfair to claim the existence of anticipation, and through the plot, we don't find the evidence of pre-treatment shifts in outcomes, therefore we will set δ to 0.

Figure E.19: Aggregate Event Study of Smartphone Adoption with No Anticipation



doi:10.6342/NTU202502893