Public Policy Research Funding Scheme

公共政策研究資助計劃

Project Number :
項目編號： 2021.A6.174.21B

Project Title :
項目名稱： Evaluation of 'Over-tourism' Phenomenon in Hong Kong Using Machine Learning and Social Media Data
以社交媒體大數據和機器學習評估香港的"過度旅遊"現象

Principal Investigator :
首席研究員： Dr LIU Xintao
劉信陶博士

Institution/Think Tank :
院校／智庫： The Hong Kong Polytechnic University
香港理工大學

Project Duration (Month):
推行期 (月) ： 12

Funding (HK$) :
總金額 (HK$)： 223,100.00

Evaluation of 'Over-tourism' Phenomenon in Hong Kong Using Machine Learning and Social Media Data

以社交媒體大數據和機器學習評估香港的"過度旅遊"現象

(Project Number: 2021.A6.174.21B)

Final Report

By

Dr LIU Xintao

The Hong Kong Polytechnic University

Acknowledgment: This research project (Project Number: 2021.A6.174.21B) is fully funded by the Public Policy Research Funding Scheme of The Government of the Hong Kong Special Administrative Region.

August 2022

Research Team

Principal Investigator

Dr. LIU Xintao
Associate Professor
Department of Land Surveying and Geo-Informatics
The Hong Kong Polytechnic University

Co-Investigators

Dr. LIN, Ming Chu
Associate Professor
School of Hotel and Tourism Management
The Hong Kong Polytechnic University

Prof. GUO, Song
Professor
Department of Computing
The Hong Kong Polytechnic University

Prof. WANG Shuaian
Professor
Department of Logistics and Maritime Studies
The Hong Kong Polytechnic University

Prof. CHEN Anthony
Professor
Department of Civil and Environmental Engineering
The Hong Kong Polytechnic University

Research Team Member

Miss TAI Kai Wing
Research Assistant
Department of Land Surveying and Geo-Informatics
The Hong Kong Polytechnic University

Executive Summary

I. Abstract

In Hong Kong, the tourism industry used to produce $328 billion visitors spending in 2018 (Hong Kong Tourism Board, 2019), which is approximately 11% of the city's GDP (Census and Statistics Department, 2019a). Tourism does boost the domestic economy. But together with the flourishing development of tourism is the negative impacts on the city, such as locals' dissatisfaction and environmental problem. If problems outweigh the benefits, the sustainability of the city will be harmed, which is also known as Over-Tourism (OT). However, there is no research specifically discussing and evaluating Hong Kong OT situation so this research fills this gap by assessing Hong Kong OT phenomenon.

The previous OT research is mainly focused on the excessive number of visitors. OT is not just counting the in-and-out number of a place. It is affected by people's perceptions. There are omissions in the evaluation of people's satisfaction and spatial interaction. This research therefore proposes a comprehensive OT index system, which integrates all these essential elements, in order to quantitively reflect the real OT situation. Furthermore, Hong Kong tourism research tends to use traditional data and methods. In order to enhance the methodology of the study, social media data and the latest data processing techniques are used in this project.

The objectives of this study are to develop an integrated database to manage spatial big data for tourism analysis; to establish indicators to quantify travel demand, satisfaction, and interaction between tourists and residents; to develop a quantitative over-tourism index; to design and implement a decision support tool to facilitate the sustainable development of tourism in Hong Kong.

OT index is composited of three indicators, which are travel demand (TD), travel satisfaction (TS) and travel centrality (TC). Twitter and Weibo data are collected. Population distribution estimation, sentiment analysis and network analysis have been performed on the data. Consequently, three indicators' results are integrated into as OT index. To help the public and the policymakers understand the OT results, an online web tool was developed to visualize the results.

The results show that OT did exist in Hong Kong and was concentrated in several districts, Wan Chai, Yau Tsim Mong, Islands and Central & Western. However, with the outbreak of coronavirus disease in 2019, the situation was mitigated by the decrease in tourists. Moreover, districts with high OT index are mainly affected by TD, while districts with low OT index are mainly affected by TS. Based on the results, five suggestions are proposed to alleviate the OT situation.

- Relieving the pressure of high travel demand districts.
- Enhancing locals' and tourists' satisfaction.
- Reviewing the transportation system and tourism-related facilities.
- Prioritizingy different tourism recommendation strategies according to the situation of each district.
- Developing a tourism app.

This research is served as the pioneer in the development of a comprehensive indicator system for evaluating the OT situation so as to inspire more research in this aspect. And it aims at helping policymakers better understand the problem and then introduce suitable strategies to tackle the problem.

II. 摘要

旅遊業是香港其中一個重要經濟產業。根據香港旅遊發展局於 2019 年發表的數據，2018 年旅客的消費總額達到三千二百八十億元。金額總值佔香港本地生產總值的十一個百分比（香港政府統計處，2019 年）。旅遊業確實可以促進本地經濟的發展，但是，伴隨著旅遊業的蓬勃發展，香港也受到負面影響，如引發本地居民不滿的情緒和環境問題。如果問題多於利益，城市的可持續性就會受到損害，這情況被稱為過度旅遊。然而，目前還沒有專門討論和評估香港過度旅遊情況的研究，因此本研究為填補此空缺而評估香港過度旅遊現象。

以往過度旅遊研究主要集中在遊客數量過多的問題上。過度旅遊不僅僅是統計一個地方的進出人數。過度旅遊情況同時會受人們的感受影響。而現時的研究在評價人們的滿意度和空間互動方面存在遺漏。本研究因而提出一個全面的過度旅遊指標體系，旨在整合所有基本要素，以量化地反映真實的過度旅遊情況。此外，香港旅遊研究大多傾向使用傳統的數據和方法。為了提升研究的方法，本項目採用了社會媒體數據和最新的數據處理技術。

總括而言，本研究的目標是開發一個綜合數據庫來管理用於旅遊分析的空間大數據；建立指標來量化旅遊需求、旅遊滿意度和遊客與居民之間的互動；建立一個反映過度旅遊情況的量化指數；設計和實施一個決策支持工具來促進香港旅遊業的可持續發展。

過度旅遊指數由三個指標組成，即旅遊需求、旅遊滿意度和旅遊中心度。本研究收集了推特和微博的數據。通過對數據進行了人口分佈估計、情感分析和網絡分析，分別得出三個指標的結果，再繼以將其整合並成為過度旅遊指數結果。為了幫助公眾和政策制定者理解過度旅遊指數的結果，本研究開發了一個在線網絡工具，將結果可視化。

結果顯示，香港確實存在過度旅遊情況，而且集中在灣仔、油尖旺、離島和中西區。但隨著 2019 冠狀病毒病的爆發，遊客的減少使情況得到了緩解。此外，高過度旅遊指數的地區主要受到旅遊需求的影響，而低過度旅遊指數的地區主要受到旅遊滿意度影響。根據研究結果，提出了五項建議幫助改善過度旅遊：

- 緩解高旅遊需求地區的壓力；
- 提高本地居民和遊客的滿意度；
- 審查交通系統和旅遊相關設施；
- 根據各區的情況分別優先考慮不同旅遊推薦戰略；
- 以及開發一個旅遊應用程序。

本研究旨在成為設立評估過度旅遊情況的綜合指標體系的先驅，以啟發更多評估過度旅遊的方法。同時，旨在幫助政策制定者更好地理解過度旅遊，然後制定合適的策略來解決這個問題。

III. Layman summary on policy implications and recommendations

In this study, five suggestions are proposed.

1. Relieving the pressure of high travel demand districts.

   Tourist spot suggestions should be evenly spatial distributed so as to avoid the high concentration of tourists in one place. Just like the picnic spots suggested by the Hong Kong Tourism Board should not only be concentrated in Yau Tsim Mong and Central & Western, but other areas should also be included. Moreover, districts with high over-tourism should be least promoted while that with low over-tourism can be highly boosted so a shift in the concentration of tourists can be achieved.

2. Enhancing locals' and tourists' satisfaction.

   By examining tourism satisfaction scores and the corresponding keywords, the satisfaction of local residents and tourists can be intentionally enhanced. To illustrate, if the district with high travel satisfaction has the most occurrence of the keyword "food", the focus of promotion in that region is on "food" and "restaurants". On the other hand, if "traffic congestion" is always found in the low travel satisfaction district, the transportation system should be revised and improved.

3. Reviewing the transportation system and tourism-related facilities.

   The transportation system between places can be revised according to the degree of travel centrality and over-tourism results. For instance, if Central & Western, one of the high over-tourism districts, has a strong connection with Sha Tin, then the transportation system connecting these two areas needs to be evaluated to ensure that the system can transport large numbers of tourists without affecting the daily lives of local residents.

4. Prioritizingy different tourism recommendation strategies according to the situation of each district.

   For the launch of tourism strategies in each district, it should be prioritized with reference of the over-tourism scores. If the district is greatly affected by travel demand, then the policy of relieving the pressure of massive tourists should be firstly planned and initiated.

5. Developing a tourism app.

   For the long-term evaluation of Hong Kong over-tourism situation, a tourism app can be developed. It should include the function of check-in, scoring, commenting, journey planning, and navigation. The app aims at improving tourists' travel experience and helping them in their time of need.

IV. 針對政策影響和建議的摘要

本研究提出了五項建議。

1. 緩解高旅遊需求區的壓力。

旅遊景點的建議應在空間上平均分佈，以避免遊客高度集中在一個地方。例如；香港旅遊發展局建議的野餐地點不應該只集中在油尖旺和中西部，其他地區也應該包括在內。此外，高過度旅遊的地區應減少推廣，而低過度旅遊的地區則可加強推廣，達致轉移遊客的集中度。

2. 提高本地居民和遊客的滿意度。

通過審查旅遊滿意度得分和相應的關鍵詞，達致目的性地提高本地居民和遊客的滿意度。舉例來說，如果旅遊滿意度高的地區，「美食」出現的次數最多，該地區推廣的重點就是「美食」和「餐廳」。另一方面，如果「交通擁塞」總是出現在旅遊滿意度低的地區，該地區的交通系統需要作出改善。

3. 審查交通系統和旅遊相關設施。

可以根據旅遊中心度和過度旅遊指數的結果來修訂地方之間的交通系統。例如，如果高過度旅遊指數的中西區與沙田有很強的聯繫，那麼連接此二地方的交通系統需要受到評估，以確保系統可以在不影響本地居民的日常生活下運輸大量遊客。

4. 根據每個地區的情況，優先考慮不同的旅遊推薦策略。

對於每個地區的旅遊推薦策略，應該參考過度旅遊指數來確定其優先次序。如果該地區受旅遊需求指標影響較大，則應首先規劃和啟動緩解大量遊客壓力的政策。

5. 開發一個旅遊應用程式。

為了長期評估香港過度旅遊的情況，可以開發一個旅遊應用程式。應用程式應該包括簽到、打分、評論、旅程規劃和導航等功能。此應用程序旨在改善遊客的旅行體驗，並能在他們需要時伸出援手。

Contents

List of Figures

List of Tables

1 Introduction

1.1 Tourism in Hong Kong

In recent years, it has been witnessed that tourism has become one of the biggest sectors contributing to the world economy as well as employment. In Hong Kong, the tourism industry used to produce $328 billion visitors spending in 2018 (Hong Kong Tourism Board, 2019), which is approximately 11% of the city's GDP (Census and Statistics Department, 2019a). This demonstrates the importance of tourism in Hong Kong. Hence, both academia and industry are eager to better understand different aspects of tourism in Hong Kong. In the early years, the overall interests were to promote tourism because it was proven that the growth of tourism can help boost the domestic economy in the short term (Jin, 2011). Some researchers made efforts to investigate the influential factors related to travel demand, such as income, travel cost, and distance (Fang Bao & Mckercher, 2008; Hiemstra & Wong, 2002). Others attempted to develop a statistical model to forecast the tourism demand (Song et al., 2012). However, the long-term benefit associated with increasing tourism is still questionable (Jin, 2011). Meanwhile, overcrowded tourists in certain areas cause new urban and social problems. The problem was out-broken when the introduction of the Individual Visitor Scheme (IVS) in 2003. Local residents expressed concerns about the disturbance of tourists, such as being harmful to the protection of cultural heritage (McKercher et al., 2005) and the natural environment (Chiu et al., 2016). The dissatisfaction of local residents was also reported as an apparent phenomenon in Hong Kong (PiuChan et al., 2018; Shen et al., 2016a). In the meantime, tourists were not satisfied with the trip compared with the expectation, partly due to the overcrowded space in Hong Kong (Song et al., 2012). These negative impacts can damage the sustainability and the growth of tourism. Novel evaluation methods are necessary to make tourism sustainable for Hong Kong's economy in the long run, with consideration of both resident and tourist satisfaction (Cheung & Li, 2019). Data-driven analysis and support tools are required to tackle the over-tourism issue to balance the economic benefit and sustainable environment.

1.2 Past and Present of Over-Tourism

A significant development in the tourism industry is observed from the 2008 economic crisis when it is regarded as an important contributor for economic recovery (Russo & Scarnato, 2017). Discounts in airline transportation and the emergence of the 'sharing economy' (e.g. Airbnb and Uber) even lower the barrier of travel activities. What comes with thriving tourism is not just the remarkable economic growth, but also a scenario in which a massive concentration of tourists in a few destinations creates an unacceptable experience on stakeholders. This scenario is widely reported in both academic papers and news articles in the term of "Over-tourism (OT)" (Phi, 2020).

A common thread in early works mainly focuses on the negative impact of OT from a capacity perspective, that is, the excessive number of tourists (Caneday & Zeiger, 1991; Forster, 1964). In these studies, extreme concentration of tourists is identified to be harmful to both urban and rural communities as well as the environment. Nevertheless, the idea of evaluating tourists is not sufficient nowadays. An alternative perspective, acceptable change framework (LAC), was introduced later with more consideration on the different attitudes of stakeholders (Mccool, 1994). Regarding financial needs, people may be more tolerant of the over-tourism impact. Till in 2000s, World Tourism Organization (UNWTO) (2018) clearly defines over-tourism as "*the impact of tourism on a destination, or parts thereof, that excessively influences perceived quality of life of citizens and/or quality of visitors' experiences in a negative way*". In addition to people's perception, quality of life and quality of experience are emphasized to be a wider scope in impact evaluation. Meanwhile, various stakeholders and components (shown in Figure 1)should be involved to evaluate OT, a complex and multi-dimensional problem (Phi, 2020).

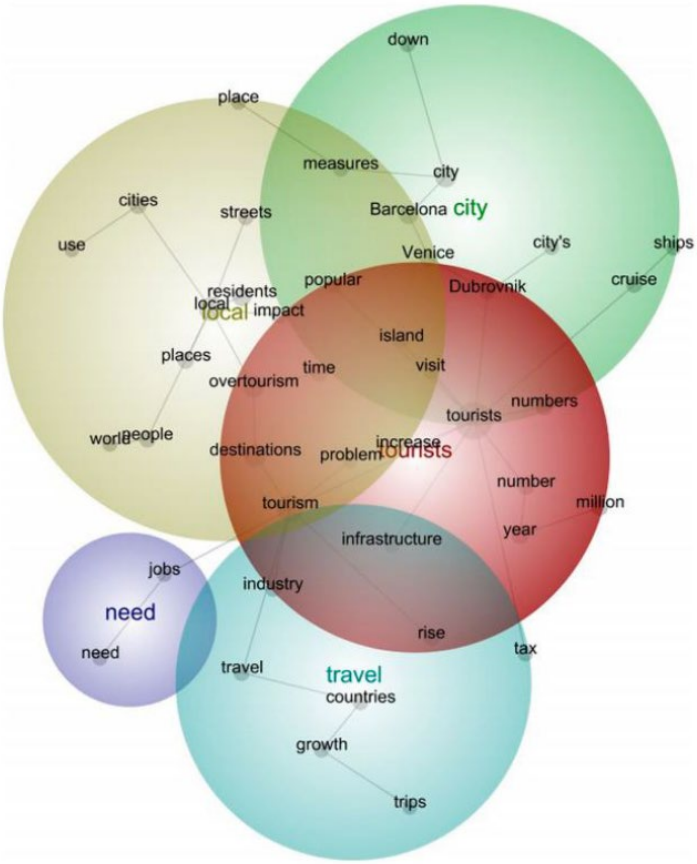


Figure 1. Important components in evaluating over-tourism impact. (Source: Phi, 2020)

Over-tourism was also happening in Hong Kong. An obvious example is that Hong Kong used to have 65 million visitor arrivals in a year, which is more than eight times its population (Census and Statistics Department, 2019b; Hong Kong Tourism Board, 2019). Massive tourism raised problems, like dissatisfactions among locals and tourists are reported (Heung & Quf, 2000; PiuChan et al., 2018). Although the situation has been relieved due to the heavy combat of COVID-19 to the global travel industry (Dutheil et al., 2020), it is forecastable that both tourism and OT will get back to their top with effective medication and vaccination for the suppression of the disease. Current is great timing and opportunity to design operational solutions to overcome Hong Kong OT problem.

However, the study of OT in Hong Kong is still very limited, mainly focusing on evaluating resident-tourist relations and attitudes (Shen et al., 2016a; PiuChan et al., 2018; Cheung & Li, 2019) but being lack of big data-driven analysis framework and tools. These studies are subject to a single aspect of over-tourism impact and are case studies using survey data, which is hard to be used as indicators for timely evaluation. To fill the gap, a new method combining high-frequency data with the existing framework is required and provides practical tools for evaluating over-tourism in Hong Kong.

1.3 Drawbacks of Previous Data and Methods

Although several previous works have contributed to the understanding of the impact of over-tourism in Hong Kong, the evaluation still faces challenges. The main drawback of the current over-tourism evaluation adopted in Hong Kong is that most of the works use interview or survey data as the data source, which fails to conduct observations with fine spatial and temporal resolution. Several representative works using traditional data and methods are presented in Table 1.

Table 1. Summary of tourism studies using traditional data and methods in Hong Kong.

Author (Year)	Title	Data	Topic	Method
Hiemstra & Wong (2002)	Factors Affecting Demand for Tourism in Hong Kong	Monthly survey	Factors related to the number of visitors	Non-linear regression model
Song et al. (2003)	Modelling and forecasting the demand for Hong Kong tourism	Annual statistical survey (Tourist arrivals and GDP)	Travel demand forecasting	Non-causal time series prediction model
Fang Bao & Mckercher (2008)	The Effect of Distance on Tourism in Hong Kong: A Comparison of Short Haul and Long Haul Visitors	Tourism and marketing survey	Behaviour and demographic relationship	Visual interpretation & indicative non-parametric statistical tests
Jin (2011)	The effects of tourism on economic growth in Hong Kong	Annual statistical survey (Visitor numbers, GDP,	Impact of tourism on economy	Vector autoregressive model

		employment)		
Song et al. (2011)	Assessing mainland Chinese tourists' satisfaction with Hong Kong using tourist satisfaction index	Face-to-face interview (N=279)	Satisfaction evaluation	Structural equation model
Shen et al. (2016b)	The sustainable tourism development in Hong Kong: An analysis of Hong Kong residents' attitude towards mainland Chinese tourist	In-depth interviews (N=38)	Residents' attitudes & sustainable tourism	Qualitative analysis using social exchange theory
PiuChan et al. (2018)	Economic and socio-cultural impacts of Mainland Chinese tourists on Hong Kong residents	Semi-structured interviews (N=10)	Influence on residents	Qualitative content analysis
Cheung & Li (2019)	Understanding visitor–resident relations in overtourism: Developing resilience for sustainable tourism	Public sentiment index and tourist report	Visitor-resident relations	Autoregressive-Distributed lag (ARDL) model

The lack of high spatial resolution of the over-tourism index makes it difficult to design customized strategies for different communities. Spatial resolution is not only important for measuring tourism intensity, but also for measuring local sentiments. Such local sentiments are an excellent practice, as it was mentioned in the LAC framework that some communities may be acceptable for a certain degree of over-tourism because of economic benefits. In terms of temporal dimension, traditional data is hard to cover long period with small time interval. For example, a finer temporal resolution will help us to the dynamic change of over-tourism metrics, which indeed is not a constant given by the tourism seasonality. The OT issue may be extreme in some periods while recovering to the normal level in other time. High frequent user-generated data can feasibly make the observation at different time frames, like by quarter, by month, by week, or even in night-time only. The flexibility of the measurement cannot be comparable with traditional data and method. Hence, related policy on over-tourism can be designed according to the fine observation.

The second drawback of the existing scheme is that survey data requires extensive labor and might be subjective. A set of new survey data need to be collected for each evaluation, whose

sample size highly depends on the project budget. It is hard to keep consistency on either sample size or sample group attributes, producing results that are relatively not comparable across cases. For measuring stakeholders' satisfaction, previous studies attempted to design survey to collect participants' opinions that will be further used to interpret human perception. This scheme is useful to allow researcher to intentionally collect certain facets of satisfaction, however, it highly depends on the skills of the survey designer. Moreover, the participants' bias is also reported due to their awareness of being interviewed. There is a gap to take advantage of 'volunteered' big data to conduct the observation on activities and sentiments as measurements of over tourism.

The third drawback of is that there is no study to quantify people's interaction as metrics in over-tourism evaluation. While interaction is an important characteristic to measure the overall importance of an entity (e.g. place) from a systematic view. Considering place-to-place interaction will allow the over-tourism evaluation model to capture not only the impact on people, but also the impact of the city as a whole inter connected system. Marketing surveys cannot provide sufficient information on interaction, while the geo-tagged big data in the proposed study will capture human mobility across space and is used to extract travel flows from place to place. The variation and disturbance of tourism activities will be captured through the new data and method.

1.4 Advantages of New Data and Method

Due to the rapid advancement of computing and internet technologies, new devices and services significantly change the way we live, and substantially, change the way of data generation. The variety of large-scale data (e.g. big data) has opened a new age for scientific research and industry applications, such as in business, management, engineering, tourism, etc. (Hashem et al., 2015; Kambatla et al., 2014). In tourism-related studies, it is not surprising that social media is one of the most widely used platforms to collect big data, because of the extensive amount of tourism-related content generated on these platforms (Del Vecchio et al., 2017; Li et al., 2018).

Social media data related to tourism, in a wider scope, include check-in locations and post contents on social networking platforms such as Twitter, or comments and ranks on review platforms such as TripAdvisor (Figure 2). Also known as user-generated content (UGC) data, the effectiveness of social media in smart tourism has received abundant discussions in recent works of literature (Del Vecchio et al., 2017; Xu et al., 2019).

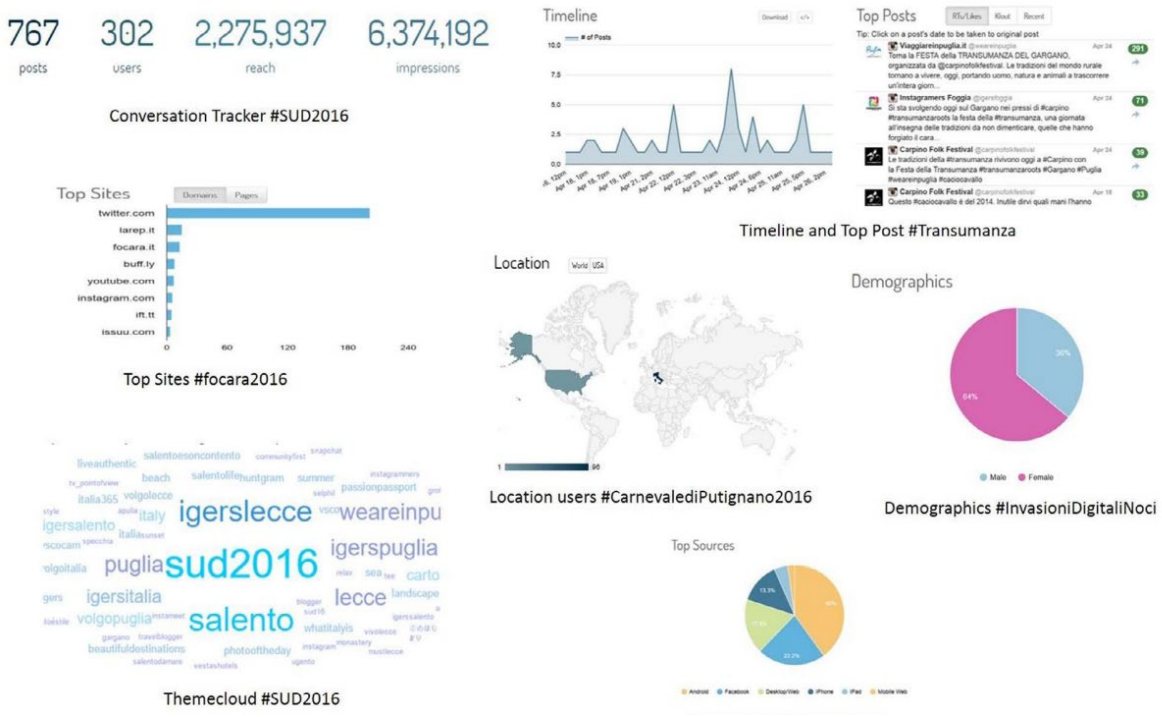


Figure 2 Social media big data for tourism. (Source: Del Vecchio et al., 2017)

To address the challenges and drawbacks of traditional data, social media big data is proposed to model over-tourism.

First, social media data can capture travel activities in high resolution of space and time dimensions. As one of the UGC data, contents and locations shared on social media are consistently generated, which means that the data volume and frequency are significantly higher than surveys. The activity locations, in the tourism context, can be used to predict travel demand (Liu et al., 2018), evaluate travel destinations (Silva et al., 2020), and understand travel

preferences among different social groups (Liu et al., 2018). The temporal variation of travel activities can be depicted in detail, which is used as signatures of different tourism sites. Furthermore, extracting spatial-temporal behavior provides an important basis to many other following tourism analysis (Xu et al., 2019). In a word, social media big data is a solid and effective lens to observe frequently updated travel activities to model over-tourism.

Second, social media data has advantages over traditional data in terms of evaluating travel experiences. Contents posted on platforms are the passive demonstration of human thoughts and feelings. Unlike an actively involved experiment environment, individuals are not aware of participating in any experiments, thus reflecting the feelings in context with less disturbances. Natural language processing and machine learning models make content analysis useful to depict various scenarios, such as traffic conditions (Pan et al., 2013) and disasters (Wang & Ye, 2018). Content analysis methods can extract topics in order to reflect travel purposes in diverse aspects and can infer sentiment so as to be used as an indicator of travel satisfaction. Therefore, social media data can help fill a gap of designing and integrating satisfaction into over-tourism assessment.

Third, social media data can be used to extract interaction between places. Although travel demand and satisfaction can be studied even using traditional data, social media data provides unique observations on spatial interaction that is often ignored in over-tourism evaluation. Individual check-in locations can be organized as space-time trajectories, which will be further aggregated as orientation destination flows among places. Using representation and metrics in network science, interaction patterns indicate the importance and connectedness of a place, which should be considered as one of the components of the over-tourism impact. The impacts would be stronger in places where spatial interactions are more significant because travel flows and emotion carried by travelers in one place is contagious for other regions through networks in the city. Network analysis is a new and novel perspective at the frontier of tourism management (Lozano & Gutiérrez, 2018). The gap of improving over tourism evaluation by introducing the important facet of urban dynamics, in specific spatial interactions, will be addressed in this project.

2 Objectives

On the basis of the critical concepts in recent studies on over-tourism, human perception, and human mobility, this project aims to develop an over-tourism index in Hong Kong. To demonstrate the practical of the index system, social media big data and the latest spatial data mining techniques are used to study Hong Kong. Further implementation of the index will also be discussed.

There are four specific objectives.

1. Develop an integrated database to manage spatial big data for tourism analysis. The database indicates the designation of structure that combines various type of spatial data.

2. Develop indicators to quantify travel demand, satisfaction, and interaction between tourists and residents using multisource social media data.

3. Develop a quantitative over-tourism index considering social and environmental impacts. The index is the combination of indicators, developed in objective 2.

4. Design and implement a decision support tool to facilitate the sustainable development of tourism in Hong Kong.

All of these have been well-achieved.

3 Research Methodology

3.1 Research Framework

Figure 3 displays the research framework of this research project. The research project consists of five parts, data collection, data cleaning (Objective 1), data analysis (Objective 2), result integration for over-tourism (OT) index (Objective 3) and web-based decision support tool development (Objective 4). Details are elaborated in the following sections.

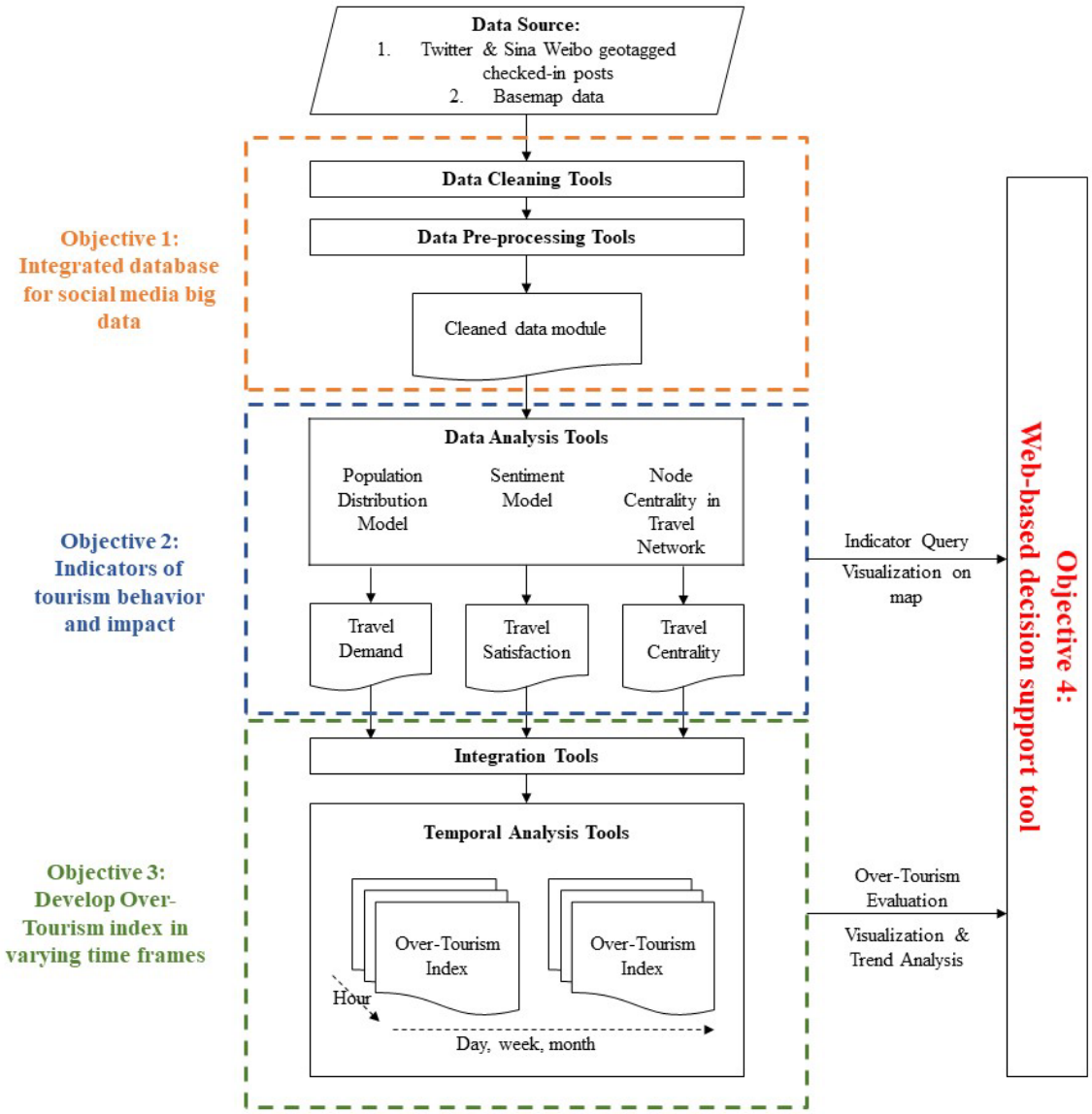


Figure 3. The research framework of the research project.

3.2 Data Sources and Tools

To extend the coverage and the diversity of users, different social media platforms' data are collected, which include Instagram, Twitter and Weibo data.

For Instagram data, it is only available in the year of 2014 to 2015. This period is unique from the period of other two data. Moreover, there is a huge disparity in the ratio of tourists to locals. The imbalance of time and identity distribution could lead to a bias in the study. Therefore, to ensure the quality and reliability of this project, Instagram data is consequently not included in the further processing and analysis while only Twitter and Weibo data are used.

For Twitter data, it is massively crawled from Twitter API by indicating the boundary of the Hong Kong region. Each post is geo-tagged, which is either a point coordinate or corner coordinates of a bounding region. The data mainly reflects the view of foreign users.

For Weibo data, it is collected through Hong Kong places keyword searching. So, the data is originally not geo-tagged. The geo-location is supplemented based on the coordinate of the keyword. Weibo data is used to capture the view of Chinese.

The summary of Twitter and Weibo data, and the reliability of the data are described in Section 4.1.

Apart from the social media data, other kinds of data are also used in different parts.

1. In the social media bot cleaning, public bot baseline datasets, cresci-2015 (Cresci et al., 2015) and cresci-2017 (Cresci et al., 2017), were used as the algorithm training dataset.
2. In the data aggregation, Hong Kong 18 districts shapefile was used (Esri China (HK), 2021).
3. In the calculation of travel demand, visitor arrivals by quarters from 2017 to 2021 (Immigration Department, 2022) and 2021 Hong Kong population (Census and Statistics Department, 2022) were included.

3.3 Objective 1: Integrated Spatial Database

Integrated Spatial Database includes part of the collected data as some of it was eliminated because of the defeats listed in section 3.3.1. Moreover, social media bots are another pollution of the data so the data belonging to those accounts were deleted from the database. Section 3.3.2 describes the technique of identifying bots. Then, section 3.3.3 discusses the classification rules of locals and tourists. The results are included in the database.

3.3.1 Data Cleaning

Raw data contains flaws, which affect further analysis, so data cleaning was taken part. The data with the following defect items are excluded from the database.

1. By user,
    a. Posts are with a null user ID.
    b. Only one post has been posted.
2. By location,
    a. The geo-tagged location is ambiguous, for example, Hong Kong, Hong Kong Island, Kowloon and New Territories.
    b. The location of the data is not within Hong Kong 18 districts region, as exemplified in Figure 4.

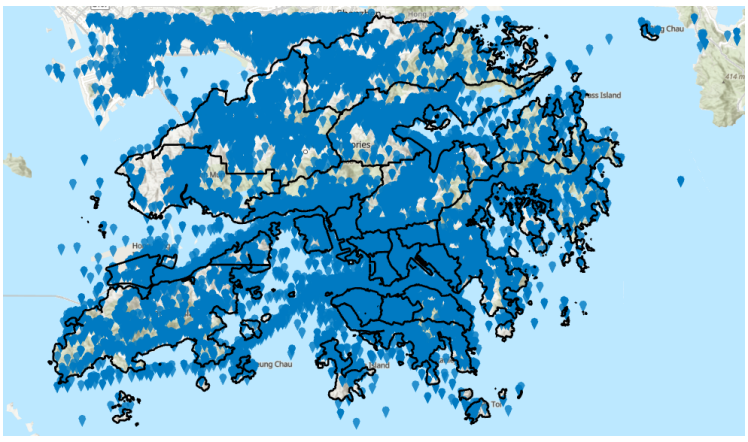


Figure 4. Visualization of the location of the raw data. Blue dots are the data while bold black lines draw the boundary of Hong Kong 18 districts. The data which is not within the boundary are excluded.

3.3.2 Social Media Bot Cleaning

Social media bots are non-human accounts on social media platforms. It could be fake accounts for spreading messages on purpose; could be business bots for posting advertisements; and could be influence bots for boosting the popularity of human accounts (Stieglitz et al., 2017). No matter what type the bots are, the data related to them could contaminate the real situation and the final results. Therefore, it is necessary to remove those data.

A social media bot detection algorithm is built to detect bot accounts and human accounts. It is based on the Random Forest, one of the machine learning techniques. By setting up rules and fitting in a suitable training dataset, the algorithm could learn the classification way and further predict the result. Considering the lack of user profile information, the rules are focusing on the content and temporal feature classes, listed in Table 2. Moreover, the datasets are available for Twitter only. The detection of other types of data is evaluated based on two obvious and common features of bots, which are the duplication of posting contents (Wang, 2010) and the high rate of posts posting in a short period of time (Stieglitz et al., 2017).

Table 2. Lists of content and temporal features for random forest bot detection algorithm.

Content Features	Temporal Features
- Post length count - Identical post count - URLs count - User mentions count - Hashtag count - Emoji count	- Posting time interval between posts count - Post in a day/ minute count

3.3.3 Local and Tourist Classification

In this study, the perspectives of locals and tourists are taken into account so a classification algorithm for the identification of locals and tourists is developed. To be reminded that, parallel-goods traders, a usual visitor type who only do parallel-goods trading, will probably be grouped into “Locals” type due to their frequent visiting to Hong Kong in a day or week or month.

The first classification rule is based on the account location, which is only applicable to Twitter data. In Twitter data, account information is collected and includes the location where the account was created. Based on this information, if the place matches the keyword of “Hong Kong” or other districts names, that account would be treated as locals. For others, it is moved to the next part.

The remaining Twitter accounts and Weibo accounts were determined by the posting time interval. The maximum posting time interval for a one-time tourist is 14 days. If the account is within this period, it is directly treated as a tourist. If it is not, the intermediate posting time interval would be considered. The minimum intermediate posting time interval for multi-time tourist is 180 days. If the posting time interval of the user can be consecutively grouped in (<14 days) and (>180 days), that user is a tourist. Otherwise, it is a local. The flow is shown in Figure 5.

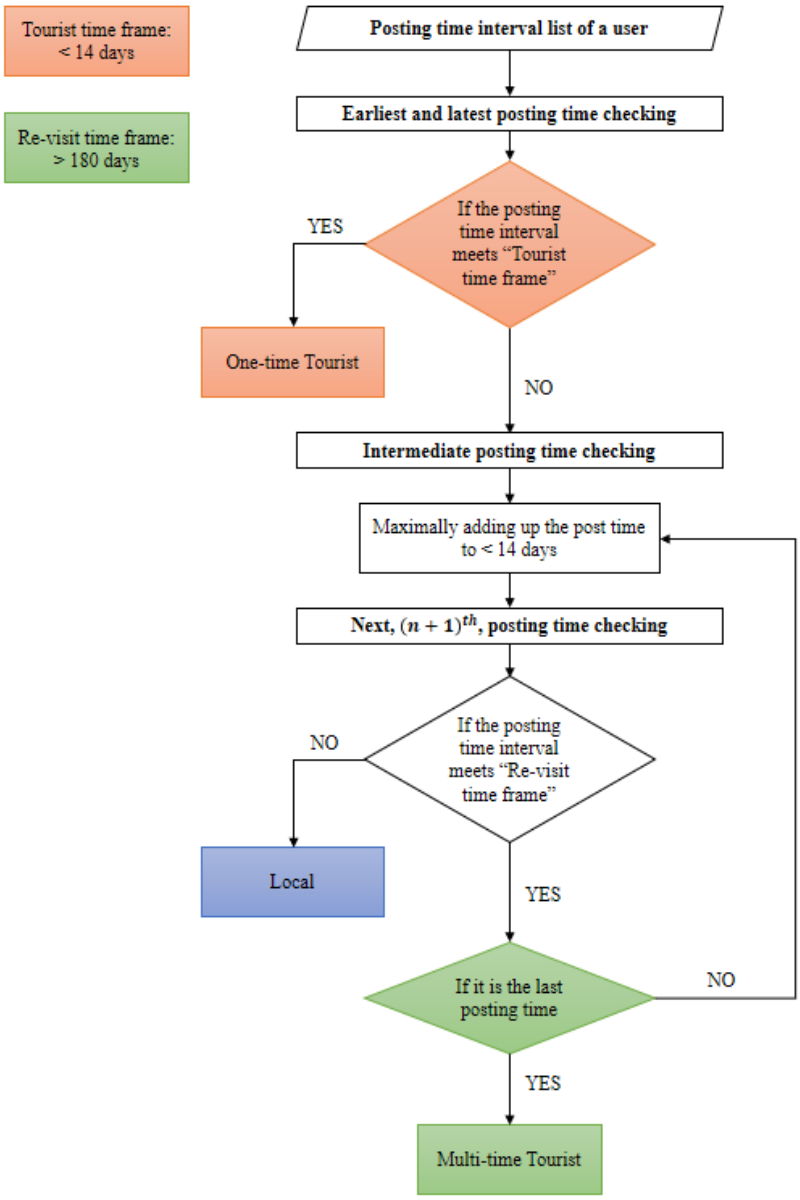


Figure 5. Flowchart of the local and tourist classification algorithm.

3.3.4 Design of Database Structure

The database is constructed for the purpose of easy retrieving and management. The data is stored in "csv" file format while the data dictionary describes and defines the meaning of the data.

3.4 Objective 2: Indicators for Over-Tourism Index

Over-Tourism (OT) index is an integrated result of three indicators, travel demand, travel satisfaction and travel centrality. As OT does not only affect tourists, but also locals, indicators include both perspectives' views. In the following sections, the model for each indicator would be discussed.

3.4.1 Travel Demand (TD)

World Tourism Organization (UNWTO) (2018) defines OT as a phenomenon of "Too many tourists" and can be measured by tourism carrying capacity and tourism congestion. The former focuses on the number of visiting tourists while the latter is about the capacity of managing tourists. To combine these two elements, a travel demand indicator is introduced, in which tourists' number represents travel demand while locals' number and area are equal to the travel supply. It is dividing the estimated visitor arrivals to the number of locals and the area of the place (in Equation 1 and 2). The quantitative value not only can present the hotspots of travel destination, but also can show the extent of the crowding of the place. The higher the value, the more popular and crowded the place. Especially, when the value exceeds 1, it may be inferred that the place is overloaded.

$$Estimated\ visitor\ arrival_{District} = \frac{Number\ of\ posts_{District}}{Total\ Number\ of\ posts_{Hong\ Kong}} \cdot Actual\ tourist\ arrival \quad (1)$$

$$Travel\ Demand\ Index_{District} = \frac{1}{2}\left(\frac{Estimated\ visitor\ arrival_{District}}{Local\ Number_{District}} + \frac{Estimated\ visitor\ arrival_{District}}{Area_{District}}\right) \quad (2)$$

3.4.2 Travel Satisfaction (TS)

Travel satisfaction is to analyze locals' and tourists' emotions based on textual information. Sentiment analysis is always a hot topic in natural language processing (NLP). There are many ways to complete the task. Sentiment scores can be calculated by dictionary-based or lexicon-

based approach. In advanced, deep learning models can be employed to classify the sentiment, like CNN (Zhang & Wallace, 2016), LSTM (Socher et al., 2013) and GNN (Ma et al., 2021). A key drawback among these methods is that it has the difficulty in fully understanding the meaning of a sentence or a paragraph due to the limit of sequential computation. To break this restriction, transformers are selected as the analysis tools in this project. The remarkable feature, self-attention, allows transformers to relate every word in the sentence and thus predict the emotion. Moreover, transformers can perform multi-language sentiment analysis if a multi-language dataset is imported for training. In this project, transformers are downloaded and used from Hugging Face (Barbieri et al., 2022; Bianchi et al., 2022; Devlin et al., 2018; Liu et al., 2019; Zhang & LeCun, 2017; Zhao et al., 2019). If the language of the post is not included in the used transformers, the text would be translated to English and classified by the corresponding transformer.

After the computation of the sentiment score of each text, the travel satisfaction index can be computed using Equation 3. It combines both locals' and tourists' sentiments.

$$Travel\ Satisfaction\ Index_{District} = \frac{1}{2}(Mean\ Sentiment\ Score_{Tourist_{District}} + Mean\ Sentiment\ Score_{Local_{District}}) \quad 3$$

3.4.3 Travel Centrality (TC)

Travel centrality concentrates on the tourists' travel trajectories in order to range the importance of places. First of all, individual travel trajectories are constructed from the sequence of social media check-in locations. To capture more reliable mobility patterns, criteria proposed by Wu et al (2014) are taken as a reference to filter out trajectories. To illustrate, consecutive locations of a tourist should have a time interval shorter than 14 days so as to present the footprints of a journey. Secondly, trajectories are aggregated to the specific spatial unit and presented in terms of nodes and links. Thirdly, PageRank centrality (Equation 4) is used to infer the relative importance of places in the whole region (Page et al., 1999; Lozano & Gutiérrez, 2018).

$$Page\ Rank\ Centrality\ (u_i) = \frac{1-d}{n} + d\sum_{j=1}^{n}\frac{PR(T_j)}{C(T_j)} \qquad 4$$

where $\boldsymbol{u_i}$ is the $\mathbf{i^{th}}$ node (the $\mathbf{i^{th}}$ district); $\boldsymbol{T_1}$ to $\boldsymbol{T_j}$ nodes are pointing to $\boldsymbol{u_i}$ node; n is the total number of nodes, d is the damping factor; $\boldsymbol{PR(T_j)}$ is the PageRank value of $\boldsymbol{T_j}$ node; $\boldsymbol{C(T_j)}$ is the number of links going out from $\boldsymbol{T_j}$ node.

3.5 Objective 3: Over-Tourism Index

Over-Tourism (OT) index consists of three indicators, travel demand (TD), travel satisfaction (TS) and travel centrality (TC). Equation 5 specifies the calculation. For TD, it is the main element in the index, ranging from 0 to unlimited. It represents the intensity of the overloading of the city. For TS and TC, they are actually capped at 1, which is acting as the sub-conditions affecting OT. So, if the OT index exceeds 1, overloading of the city is detected. To be enhanced, different weightings can be assigned. Hence, time-wise strategies, such as by days, months or years, have to be determined during the calculation.

$$Over-Tourism\ Index = \frac{1}{3}(Travel\ Demand + Travel\ Satisfaction + Travel\ Centrality) \quad 5$$

OT index results would be used to study the intensity and the change of Hong Kong OT situation and thus helping strategic decisions on tourism policy in Hong Kong.

3.6 Objective 4: Web-based Decision Support Tool

The main purpose of this web-based decision support tool is to spatially and temporally show the OT index results, so as to provide the evidence for the evaluation of Hong Kong OT situation and impact. Maps and layers to spatially present the results of TD, TS, TC and OT while the time slider must be included to help control the visualization of temporal change. Therefore, tourism spatial-temporal distribution of the OT can be observed.

4 Research Results

4.1 General Information

In this project, Hong Kong is selected as the study area. It is too vague to evaluate the situation at a city level. So, district is chosen as the spatial unit, and which total of 18 districts are included. The spatial distribution and boundary of districts is shown in Figure 6.

Figure 6. Spatial distribution and boundary of Hong Kong 18 districts.

To cover the period of COVID-19 pre-pandemic and during-pandemic period, data are crawled from 2017 to 2021.

According to the spatial and temporal setting, Twitter and Weibo data were collected. Raw data information is summarized in Table 3. However, not all the data can be fitted into the analysis. Data aggregation to the spatial unit and data cleaning as described in section 3.3 were performed. Additionally, social media bot cleaning was done. Examples of the bot accounts' contents are shown in Table 4. Those accounts not only shared meaningless posts, like User 1 and User 4, but also advertised businesses, like User 2 and User 3, which do not help assess travel experience nor people's lives. At last, only human user accounts are left while they are classified into locals and tourists. Table 5 presents the summary of the final cleaned data.

Table 3. Summary of the raw data.

Data Type	**Data Period (YYYY-MM-DD)**		**Number of Posts**	**Number of User Accounts**
	From	**To**		
Twitter Data	2017-01-01	2021-12-31	3456816	156930
Weibo Data	2018-01-01	2018-12-31	332417	157793

Table 4. Examples of social media bot accounts.

		Post	**Posting Time**
Twitter	User 1	Course #hashtag1 URL1	2018-07-12 23:56:22
		Crush #hashtag1 URL2	2018-07-12 23:55:51
		Congratulations #hashtag1 URL3	2018-07-12 23:55:17
	User 2	UKS🆕全新升級🆙電鍍粉金色 $300	2021-08-13 11:32:55
		`預告≋vapemohoⓂ黃銅雕刻✨限量版.....	2021-08-06 12:32:36
		日本與樂JP最新味🆕泰清椰皇水□	2021-07-31 13:44:19
Weibo	User 3	【请注意，万圣元素通缉令发布啦】……	2018-10-26 19:05:00
		#好玩情报局##万圣节#.....	2018-10-23 21:06:00
		你也被乌镇戏剧节刷屏了吗？……	2018-10-22 17:23:00
	User 4	【其实你不懂我的心（童安格）】……	2018-06-08 04:56:00
		【又见炊烟（邓丽君）】……	2018-06-08 04:55:00
		【飘落 这里的黎明静悄悄 主题曲】……	2018-06-08 04:55:00

Table 5. Summary of the cleaned data.

	Number of Posts	Number of User Accounts
Locals	1597645	31058
Tourists	94873	17319
Human Users	1692518	48377

According to the post contents' language (Figure 7), most of the tourists, more than 50%, are using Chinese. It is followed by English (38%). Then, for around 7% of tourists, their main languages are Japanese, Indonesian, French, Russian and Other. To validate the percentage of tourist distribution, the actual percentage of tourist arrivals by market (Figure 8) is taking as reference baseline, with the assumption that Australian, Singapore, United Kingdom and USA are grouped in English class while Mainland China and Taiwan are grouped in the Chinese class. There are large differences in the Chinese class and the English class. The main reason may due to the lack of Weibo data. This caused the decrement in the Chinese data occupancy rate in the dataset and therefore has increased the ratio of other tourists. Despite the discrepancy of the tourists' distribution in the percentage, Chinese remains getting the top occupation; English is on its second place; and Japanese, Indonesian, French and Russian come at the 3rd, 4th, 5th and 6th position. The order is approximately the same. As a result, the data is still reliable in a certain extent.

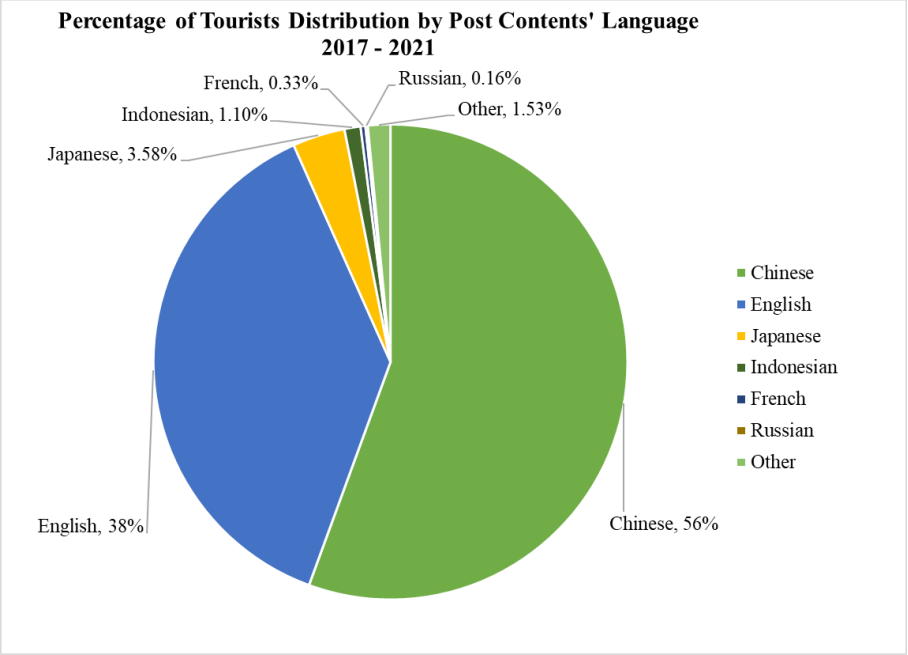


Figure 7. Percentage of Tourists Distribution by Post Contents' Language

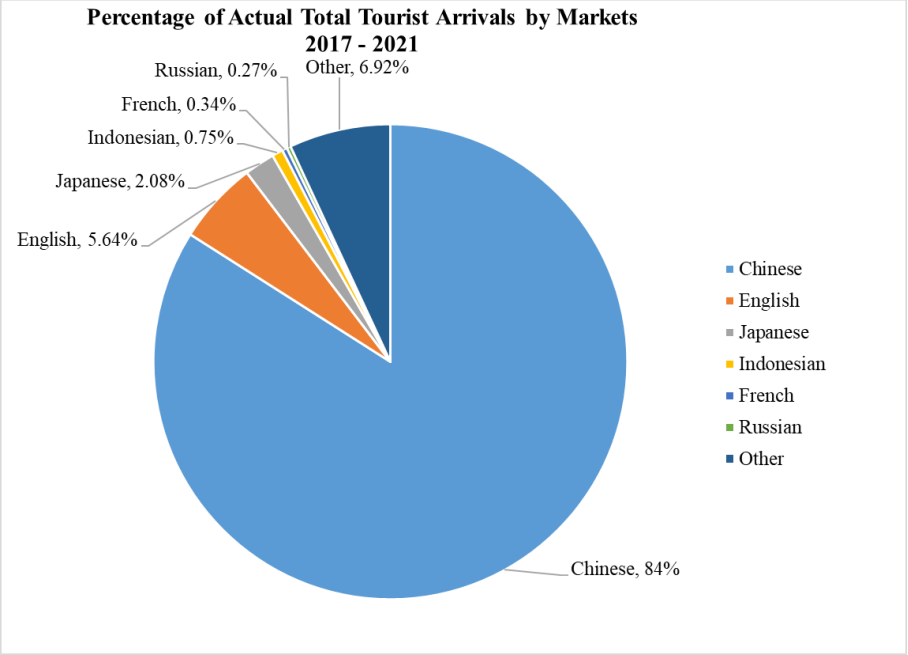


Figure 8. Percentage of Actual Total Tourist Arrivals by Markets.

4.2 Over-Tourism Index

In this section, Over-Tourism (OT) results would be discussed. Each period is in 6 months while the starting month is used as the representative term, for example, 2017/01 means the period from 2017 Jan to 2017 June, but only four months for the period 2021 Jul to 2021 Oct.

4.2.1 Overall OT Results

Figure 9 shows the overall trend of the Hong Kong Over-Tourism (OT) index. From the period of 2017/01 to 2019/01, the OT index remains over 1. After that, the index keeps falling until 2020/07. Then, the OT is holding at around 0.18. These computed results match the real trend. Before the COVID-19 pandemic, Hong Kong was a well-known tourist city with tens of millions of tourist arrivals per year. However, due to the travel restriction policies during the pandemic, the amount of traveling drops tens of thousands. The situation has not been recovered yet. Therefore, the proposed OT index model and indicators can reflect Hong Kong's tourism status in a certain extent. Moreover, it is noticed that the index broke the wall of 1 in the pre-pandemic period. This can deduce that Hong Kong kept overloading in handling tourists. OT was actually happening.

Figure 9. Hong Kong Over-Tourism Index Trend.

Figure 10 investigates the trend at the district level. Obviously, Wan Chai (dark blue), Yau Tsim Mong (orange), Islands (dark green) and Central & Western (light green) were in excess load until 2020/01. As stated, OT index can only break 1 when travel demand (TD) is overloading. Those districts with over 1 mean that the number of tourist arrivals of that place exceeds both the locals' number and the district area. The results are reasonable. For Wan Chai and Central & Western, these districts have numerous east-meets-west, modern and traditional buildings, which are the best visiting places; for Yau Tsim Mong, well-known shopping malls and restaurants attract tourists' attention; for Island, the airport in there is the best check-in location. Therefore, the results are in the expectation and this can be proof of the reliability of the proposed OT index model.

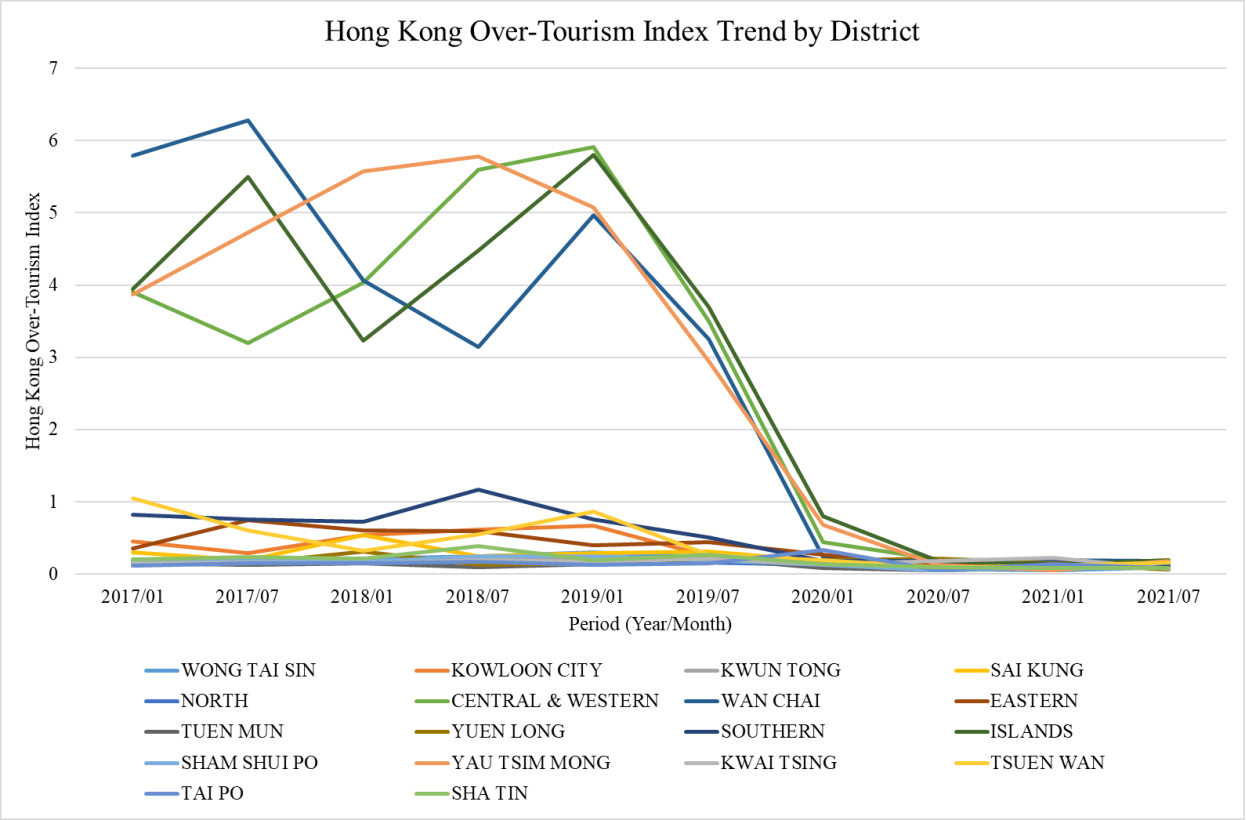


Figure 10. Hong Kong Over-Tourism Trend Index by District.

4.2.2 Travel Demand (TD) Results

Travel Demand (TD) is the divided results of the number of tourists to the number of locals and the district area, refers to Equation 2. The overall spatial and temporal change of TD is captured in Figure 12 and Figure 13. Legends (Figure 11) help the interpretation of maps.

TD results are plotted on top of the over-tourism (OT) index. It is in rhombus shape. Its color represents the level of the tourists/area and that of tourists/locals among all districts while its shape indicates the TD value. The larger the shape, the higher the TD. The darker the orange, the higher the value of tourists/locals. The darker the blue, the higher the value of tourists/area.

From the period of 2017/01 to 2019/07, high TD values are obviously spotted in Islands, Yau Tsim Mong, Central & Western, and Wan Chai. Moreover, these districts are in dark brown color, which means that they are having a comparatively high value of tourists/area and tourists/locals. To deeply investigate the two sub-indexes, tourists/locals places a more significant role on the TD value. In Figure 14, the above-mentioned districts have values over 15. Every local faces 15 tourists. This increases the chances of conflict between locals and tourists. On the other hand, the intensity of tourists/area is less (in Figure 15). Yau Tsim Mong is the only district with over 1. Every tourist has less than 1 m^2 of space to move around. This is solid proof of the overcrowding phenomenon.

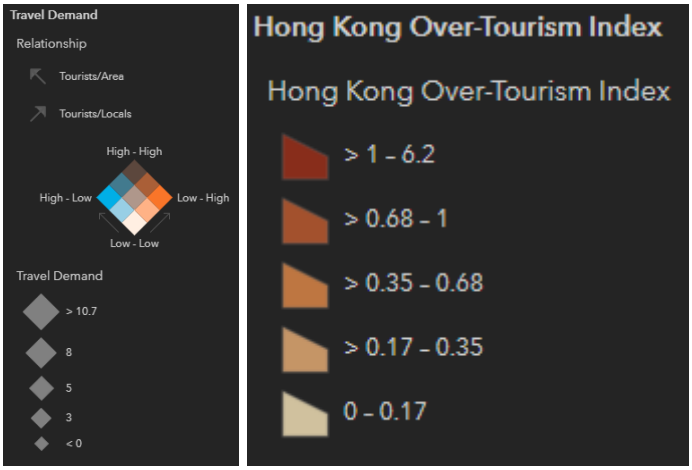


Figure 11. Legend for Spatial and Temporal Change of Travel Demand.

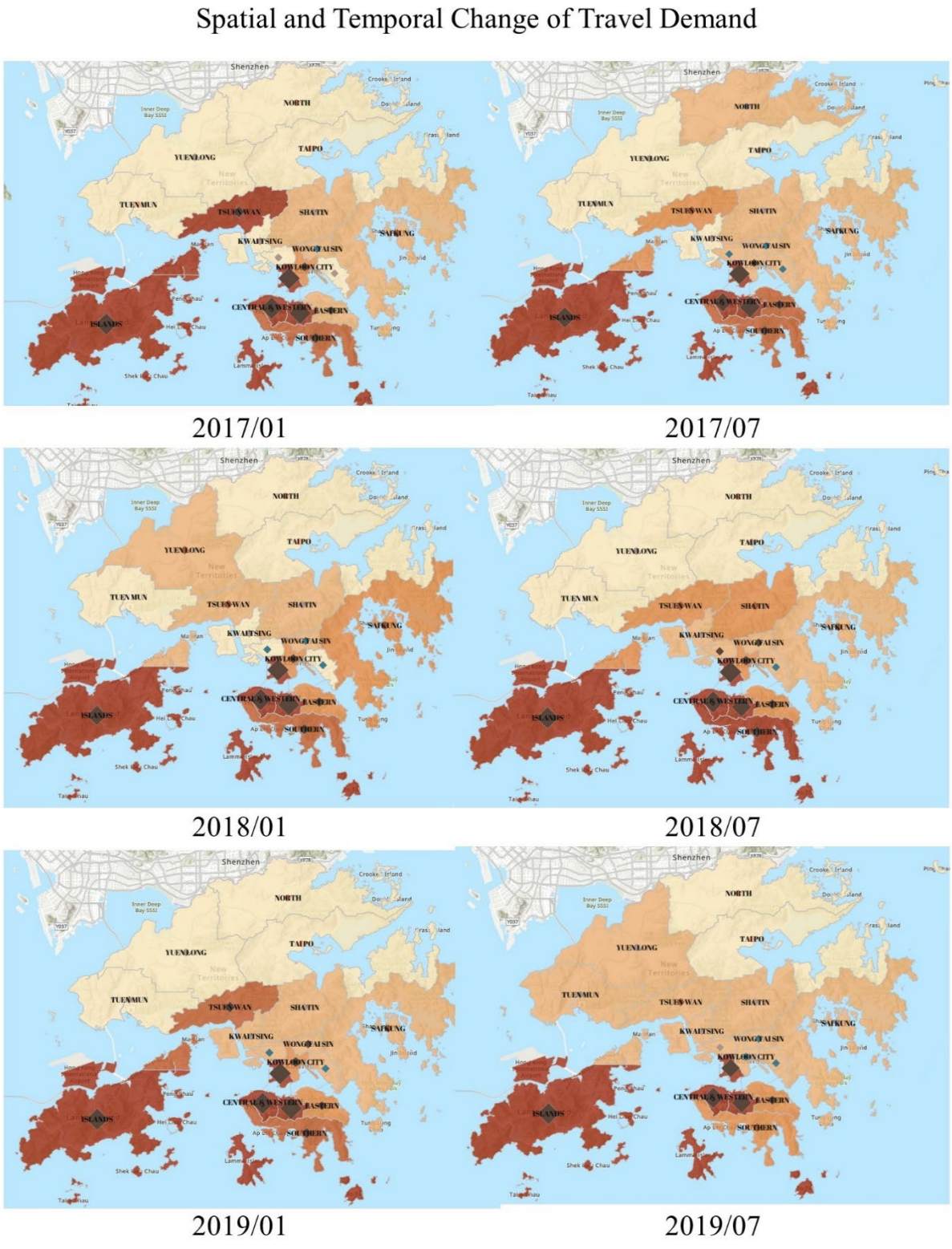


Figure 12. Spatial and Temporal Change of Travel Demand.

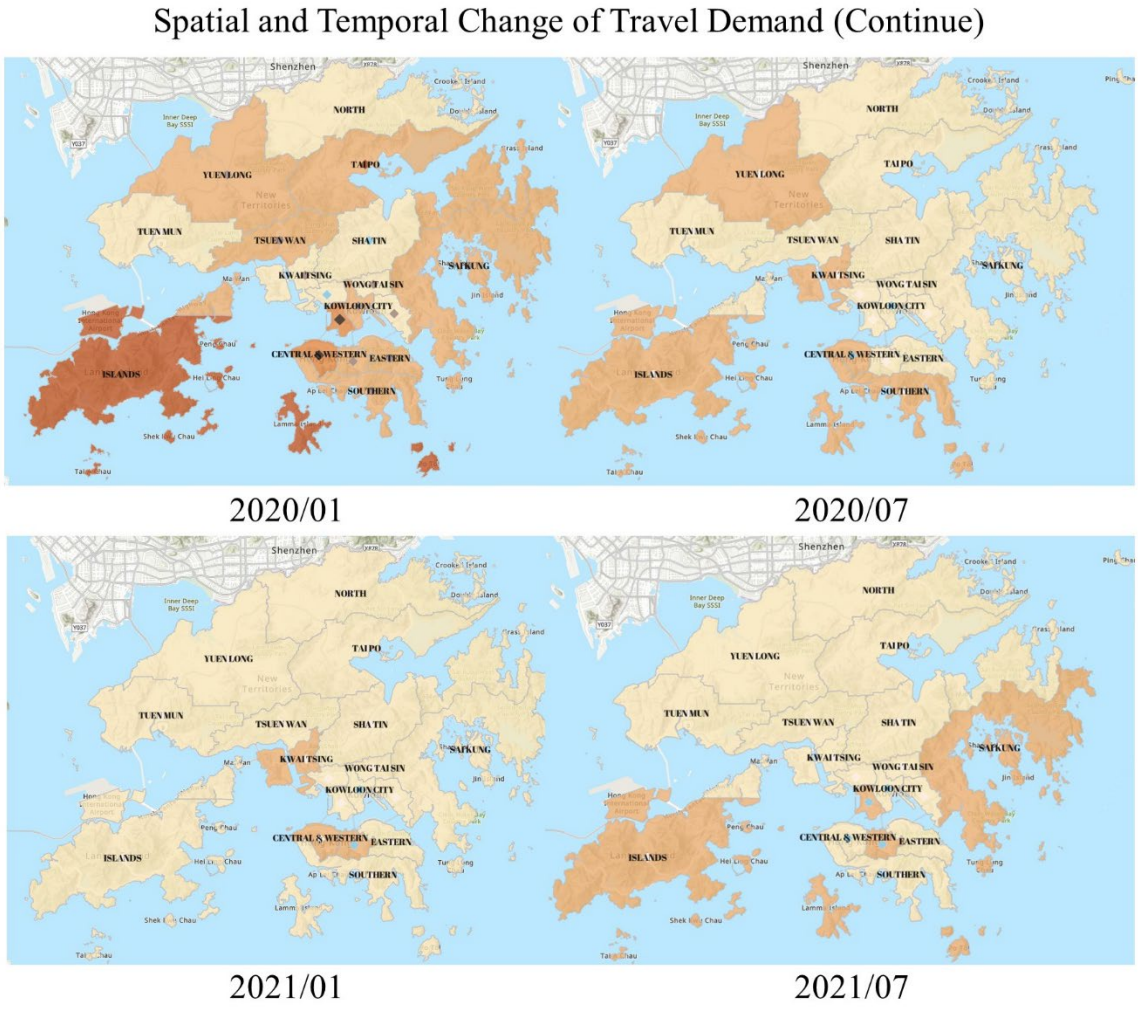
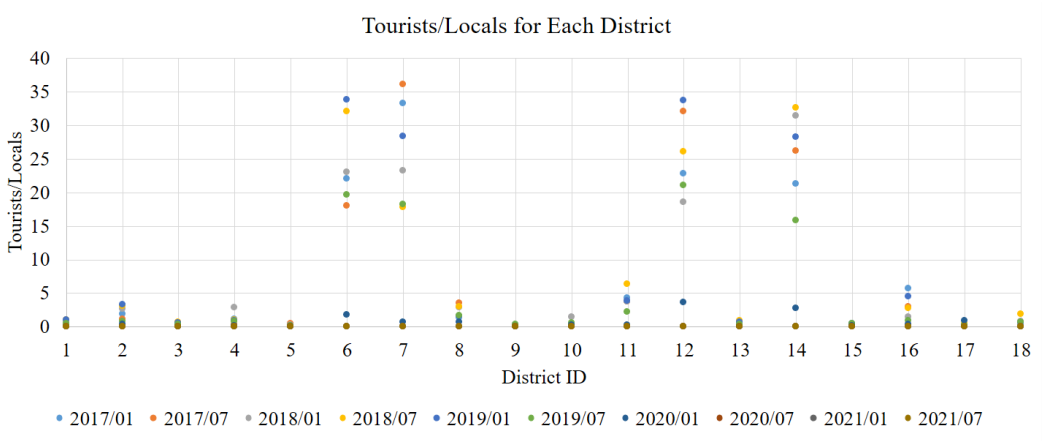
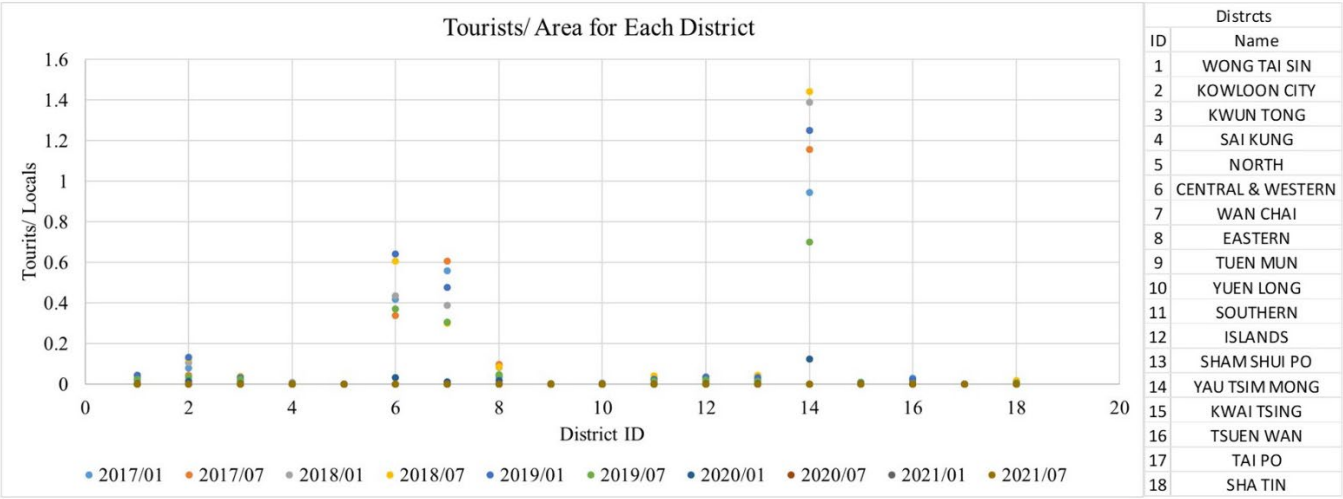


Figure 13. Spatial and Temporal Change of Travel Demand (Continue).



District ID	District Name
1	WONG TAI SIN
2	KOWLOON CITY
3	KWUN TONG
4	SAI KUNG
5	NORTH
6	CENTRAL & WESTERN
7	WAN CHAI
8	EASTERN
9	TUEN MUN
10	YUEN LONG
11	SOUTHERN
12	ISLANDS
13	SHAM SHUI PO
14	YAU TSIM MONG
15	KWAI TSING
16	TSUEN WAN
17	TAI PO
18	SHA TIN

Figure 14. Tourists/ Locals for districts.



Distrcts	
ID	Name
1	WONG TAI SIN
2	KOWLOON CITY
3	KWUN TONG
4	SAI KUNG
5	NORTH
6	CENTRAL & WESTERN
7	WAN CHAI
8	EASTERN
9	TUEN MUN
10	YUEN LONG
11	SOUTHERN
12	ISLANDS
13	SHAM SHUI PO
14	YAU TSIM MONG
15	KWAI TSING
16	TSUEN WAN
17	TAI PO
18	SHA TIN

Figure 15. Tourists/ Area for districts.

4.2.3 Travel Satisfaction (TS) Results

Travel Satisfaction (TS) is the average results of locals' sentiment and tourists' sentiment, refers to Equation 3. The overall spatial and temporal change of TS is captured in Figure 17 and Figure 18. Legends (Figure 16) help the interpretation of maps.

TS results are plotted on top of the over-tourism (OT) index. It is in square shape Its color represents the level of locals' sentiment and that of tourists' sentiment among all districts while its shape indicates the TS value. The larger the shape, the higher the TS. The darker the blue, the more positive the tourists' sentiment. The darker the pink, the more positive the locals' sentiment

Reviewing all years, the values are constantly at a steady level, around 0.7. This means that both locals and tourists generally feel positive. Moreover, there is no fixed pattern regarding the sentiment score. For example, high OT districts, Yau Tsim Mong, Central & Western, and Wan Chai, had comparative high sentiment scores in 2018, but the scores dropped in other years. The ranking of the scores always takes turns among all districts. So the more concerning issue is to investigate the factors that led to the scores.

To study the factors, word clouds are plotted. Firstly, looking into the word cloud of negative locals' sentiment (Figure 19), keywords are mostly about daily lives, such as "people", "time", "工作 (work)", "股份 (stock)" and "有限公司 (Limited Company)". It does not highly relate to tourism. But in the positive view (Figure 20), some tourists places, like "星光大道 (Avenue of Stars)", "迪士尼 (Disneyland)" and "濕地公園 (Hong Kong Wetland Park), appear. This may infer that locals also feel positive about these places. Although specific locations are spotted, the reasons causing this feeling can hardly be identified. It is because many circumstances can lead to this status, which can be not related to the tourism issue. Just like, "happy birthday" can be a reason why a person feels "great" and "Disneyland" is "amazing". Despite the unknown factors, at least some tourist attractions can be noticed from the perspective of the locals.

Then, in tourists' view, no matter it is negative (Figure 21) or positive (Figure 22), more tourists spots are mentioned, including but not limited to "銅鑼灣 (Causeway Bay)", "迪士尼 (Disneyland)", "尖沙咀 (Tsim Sha Tsui)". Factors can be deduced from keywords. Taking negative words as an example, "people" may mean the problem of crowding and "地鐵 (subway)" can be transportation issue. On the other hand, "beautiful" can be a compliment to the scenery of Hong Kong, and "好吃 (yummy)" implies the great of Hong Kong food or restaurant. These can be hints of tracing the back to the reasons. Thus, tourism improvements or recommendations can be further developed based on tourists' responses.

So TS not only can be used as an index, the words behind the numbers can be helpful for OT investigation.

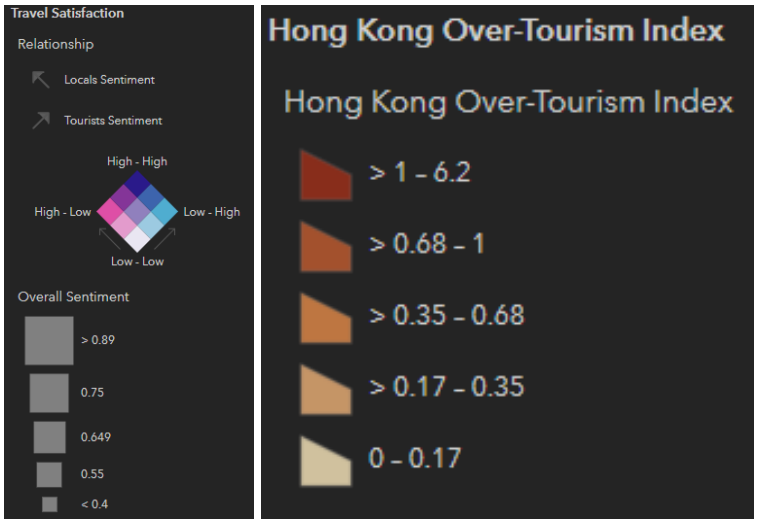


Figure 16. Legend for Spatial and Temporal Change of Travel Satisfaction.

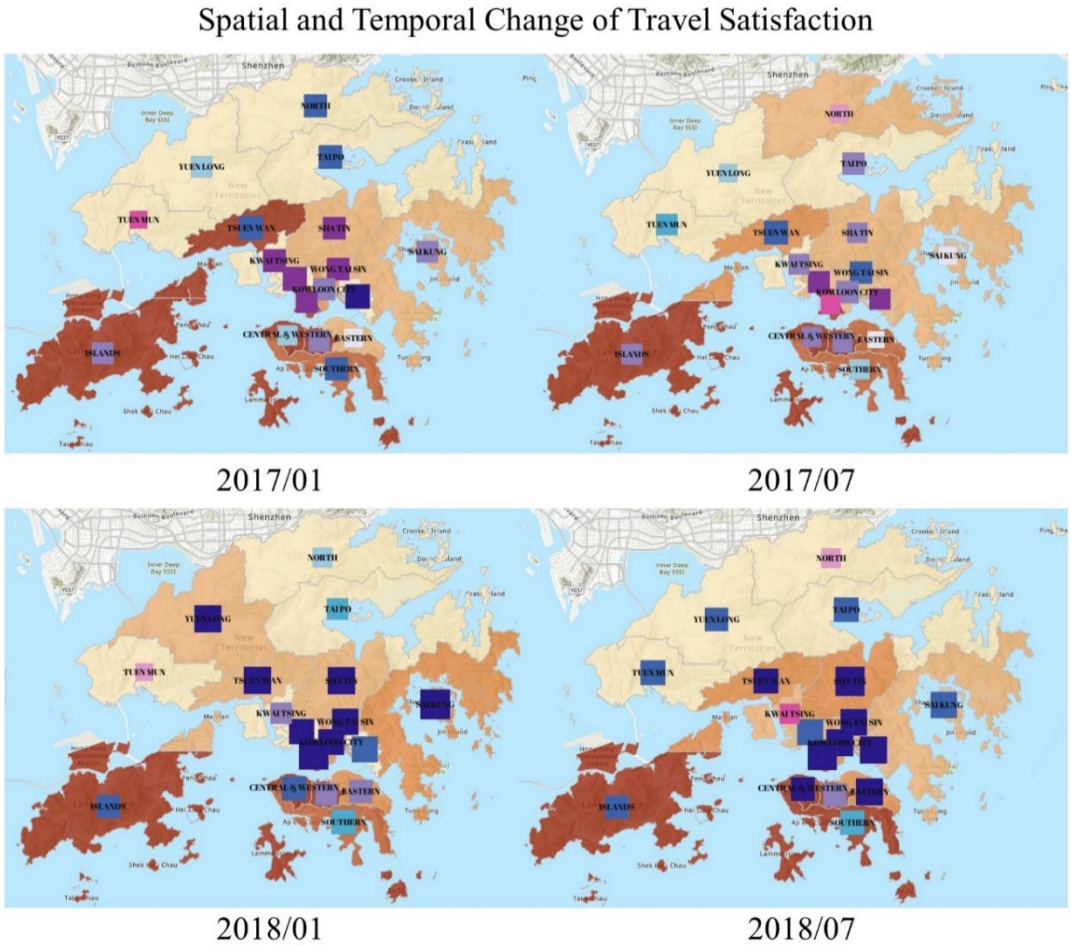


Figure 17. Spatial and Temporal Change of Travel Satisfaction.

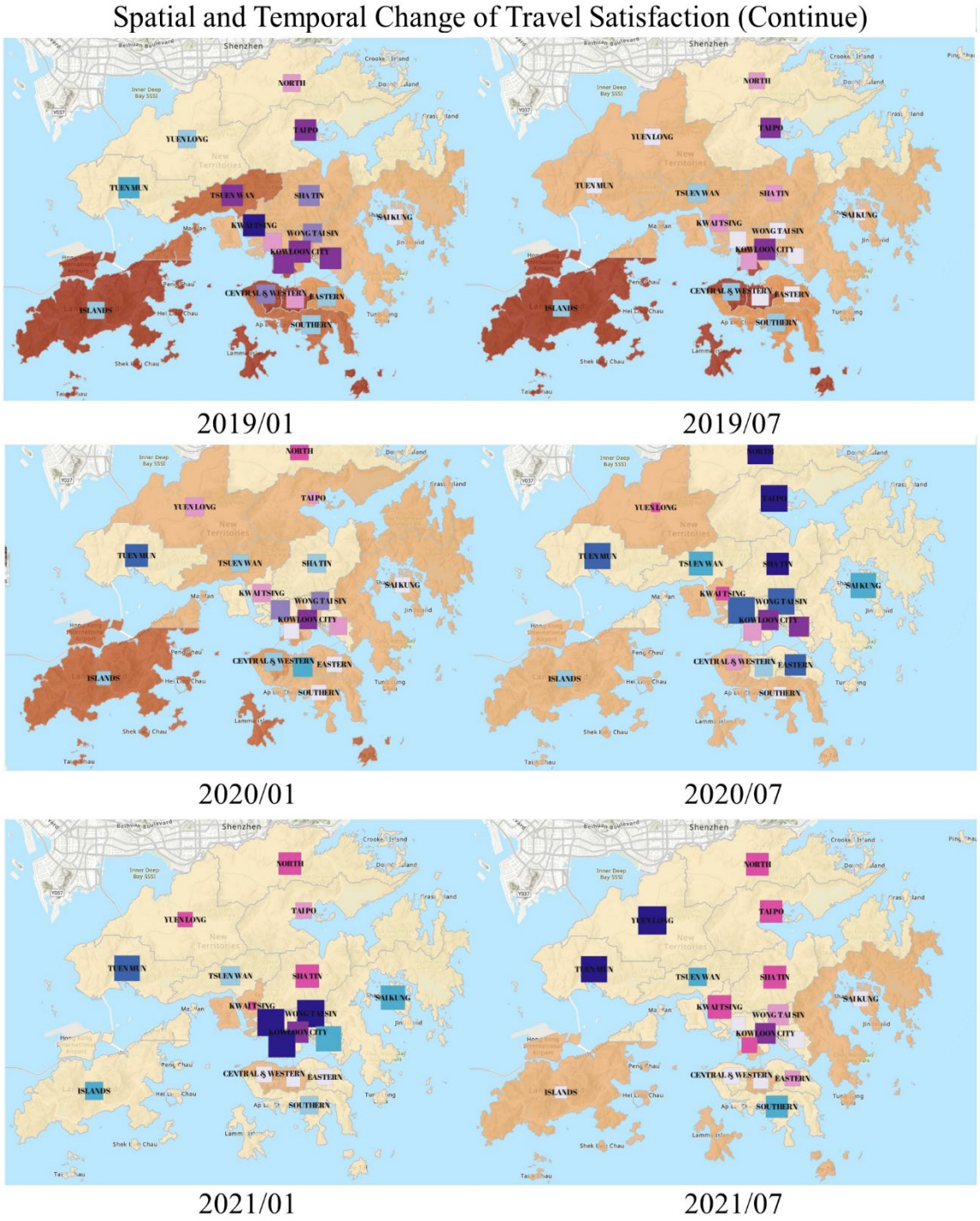


Figure 18. Spatial and Temporal Change of Travel Satisfaction (Continue).

Figure 19. WordCloud of negative locals' sentiment (Top: English posts; below: Chinese posts).

Figure 20. WordCloud of positive locals' sentiment (Top: English posts; below: Chinese posts).

Figure 21. WordCloud of negative tourists' sentiment (Top: English posts; below: Chinese posts).

Figure 22. WordCloud of positive tourists' sentiment (Top: English posts; below: Chinese posts).

4.2.4 Travel Centrality (TC) Results

Travel Centrality (TC) is the PageRank probability of a tourist to one place. The overall spatial and temporal change of TC is captured in Figure 24 and Figure 25. Legends (Figure 23) help the interpretation of maps.

TC results are plotted on top of the over-tourism (OT) index while black arrow lines indicating the direction from place to place are on its top. TC is in circular shape Its shape indicates the TC value. The larger the shape, the higher the TC. For black arrow lines, the darker the color, the more people traveled in that direction.

Centrality mostly can be used as the indicator of importance. So in TC, the higher the values, the more important the place. Again, Yau Tsim Mong, Central & Western, and Wan Chai, which have high OT index, have high TC values from the period of 2017/01 to 2019/07. Traveling is a mobility activity. People move around from place to place. So it is not surprising that OT places have the most number of arriving or departing movements. For other districts, fewer tourists traveled from/ to so the TC would comparatively be low. As TC is a relative result, the importance can be high even when it is in a during-pandemic period, as long as there is travelling.

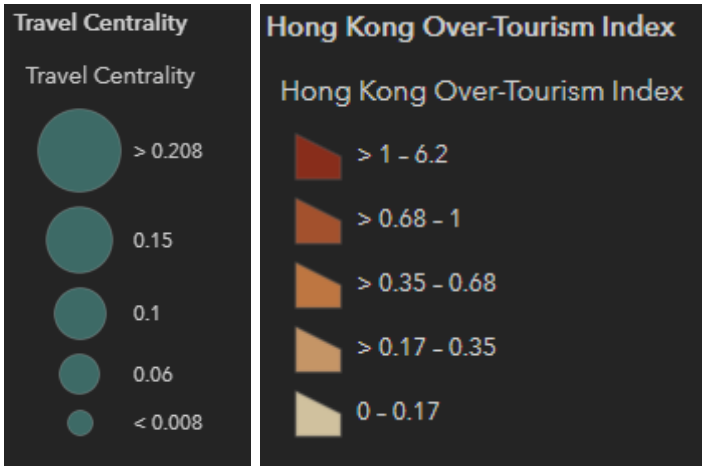


Figure 23. Legend for Spatial and Temporal Change of Travel Centrality.

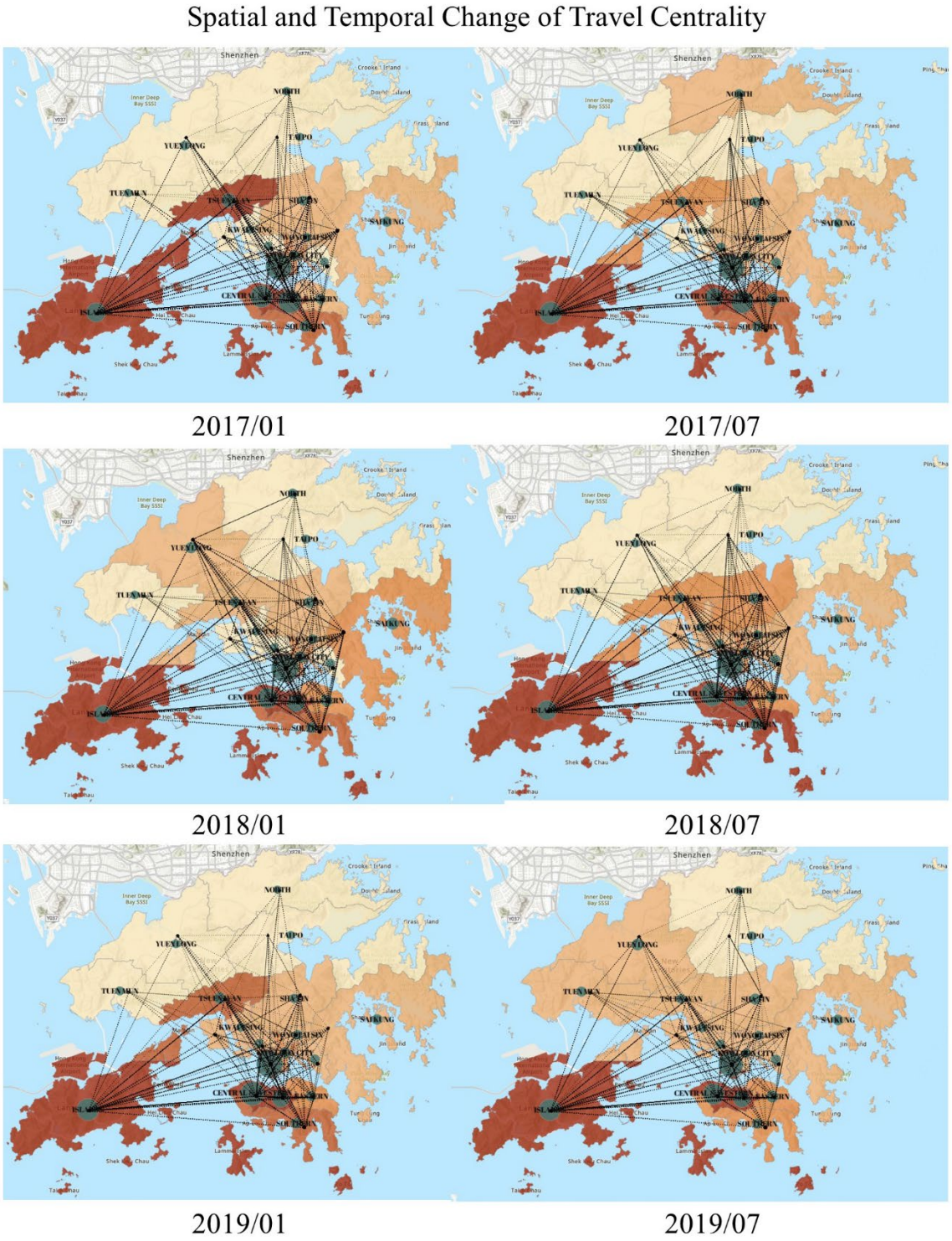


Figure 24. Spatial and Temporal Change of Travel Centrality.

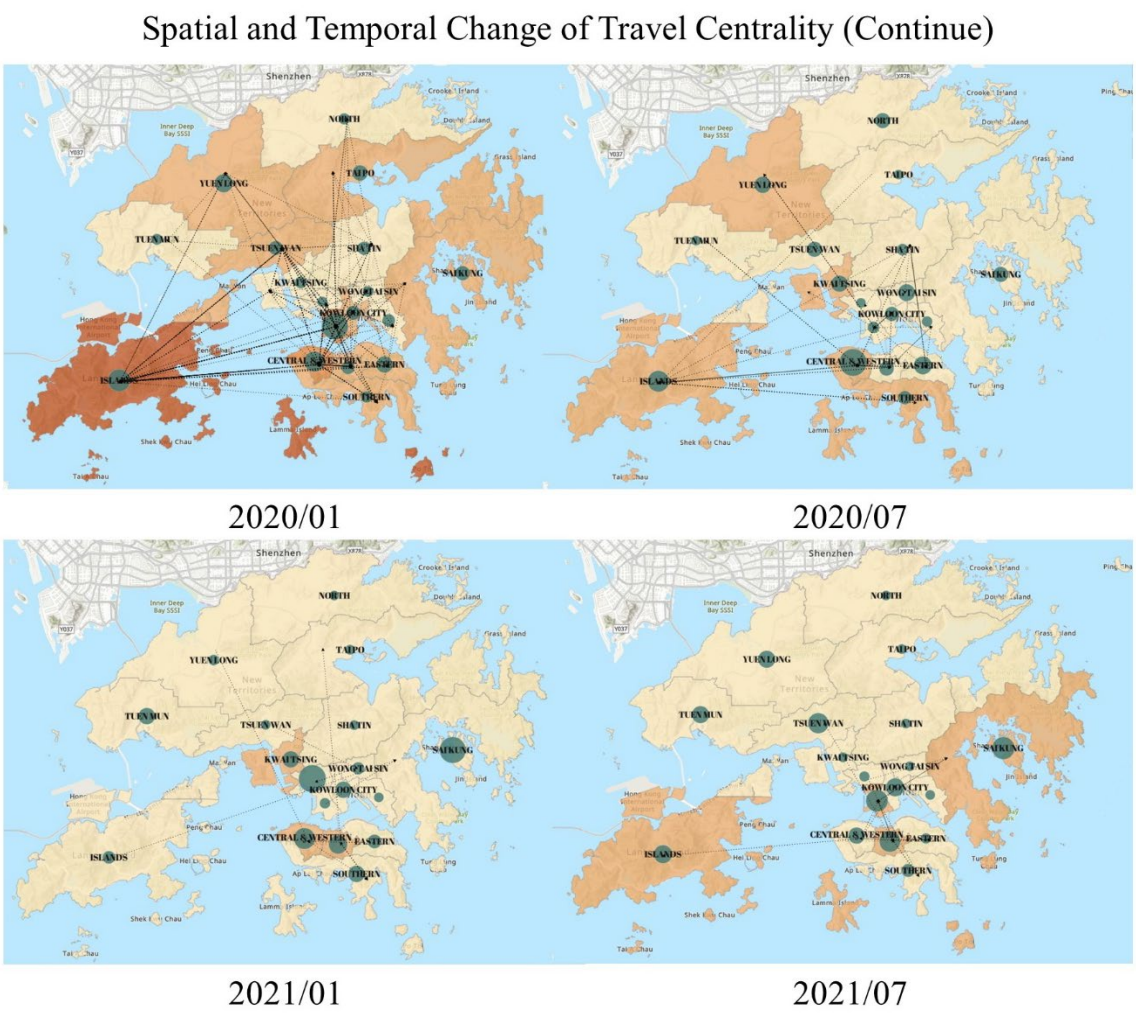


Figure 25. Spatial and Temporal Change of Travel Centrality (Continue).

4.3 Web-based Decision-Support Tool

Link: https://arcg.is/0mzHyH

The web-based decision-support tool is developed to spatially and temporally visualize the OT results. It takes the benefit of Esri ArcGIS tools to construct the platform. The overview of the tool is shown in Figure 26.

Six layers are included in the map, which is “Travel Network”, “Travel Centrality”, “Travel Satisfaction”, “Travel Demand”, “Predominance of Indicators” and “Hong Kong Over-Tourism Index”.

1. For the travel network, black arrow dash lines represent tourists’ travel trajectories from district to district. The intensity of the color of the line infers the number of trajectories. The darker the color, the more tourists traveled in that direction. Meanwhile, the centrality of the place can be deduced from it together with the number of arrows meeting at there. Figure 27 shows an example of the travel network pattern. This is usually active with TC layer.
2. For Travel Demand (Figure 28), Travel Centrality (Figure 29) and Travel Satisfaction (Figure 30), it shows the results. It can be supplementary information in analyzing OT index.
3. For the Predominance of Indicators, it shows the predominant result of three indicators (travel demand, travel satisfaction and travel centrality) in the OT result. To illustrate, in Figure 31, those dark green rhombuses are indicating that OT result of that district is mainly affected by travel demand in a large extent while those light green rhombuses are also prominently affected but just in a slight extent. The blue rhombuses are sharing the same explanation logic but in the travel satisfaction category.

4. For Hong Kong Over-Tourism Index, it displays the OT results by district. Totally five categories are used to divide the level of the OT. The darker the red color, the more severe the OT situation. Figure 32 exemplifies that Hong Kong Island, Islands and Yau Tsim Mong are having the most severe OT situation; Tuen Mun and Tai Po have the least influence from OT; and other districts are in between.

To sum up, this web-based decision-support tool is to spatially and temporally visualize the Hong Kong OT index and situation, aiming at helping to make decisions on how to develop tourism in a more sustainable manner.

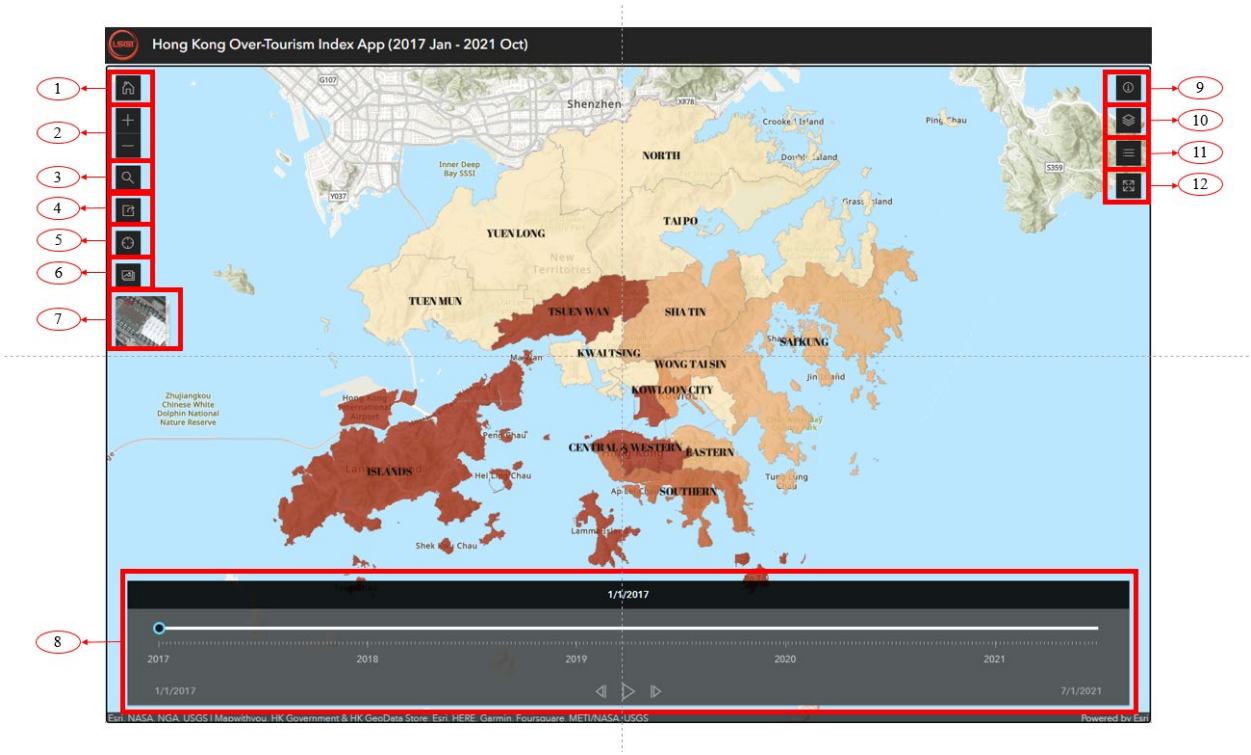


Figure 26. Overview of the web-based decision support tool. Red boxes are the available widgets for different functions. The description of the functions can be checked in the Table 6.

Table 6. Summary of the functional widgets with its description.

	Tools	Description
1	Home button	Returning the map to its default view extent.

2	Zoom controls	Zooming the map to different levels.
3	Search	Searching locations on the map.
4	Social sharing	Sharing the tool using a link or social media.
5	Find current location button	Presenting the current device location on the map.
6	Screenshot	Taking a screenshot of the map at present view.
7	Basemap toggle	Toggling the basemap in between imagery map and topographic map.
8	Time slider	Controlling the visualization of data in temporal.
9	Introduction	Showing the brief description of this tool.
10	Layer list	Displaying the list of map layers with the function of turing on or off the layer in the map.
11	Legend	Showing the legend of the present visible layers.
12	Full screen	Displaying the tool at fullscreen.

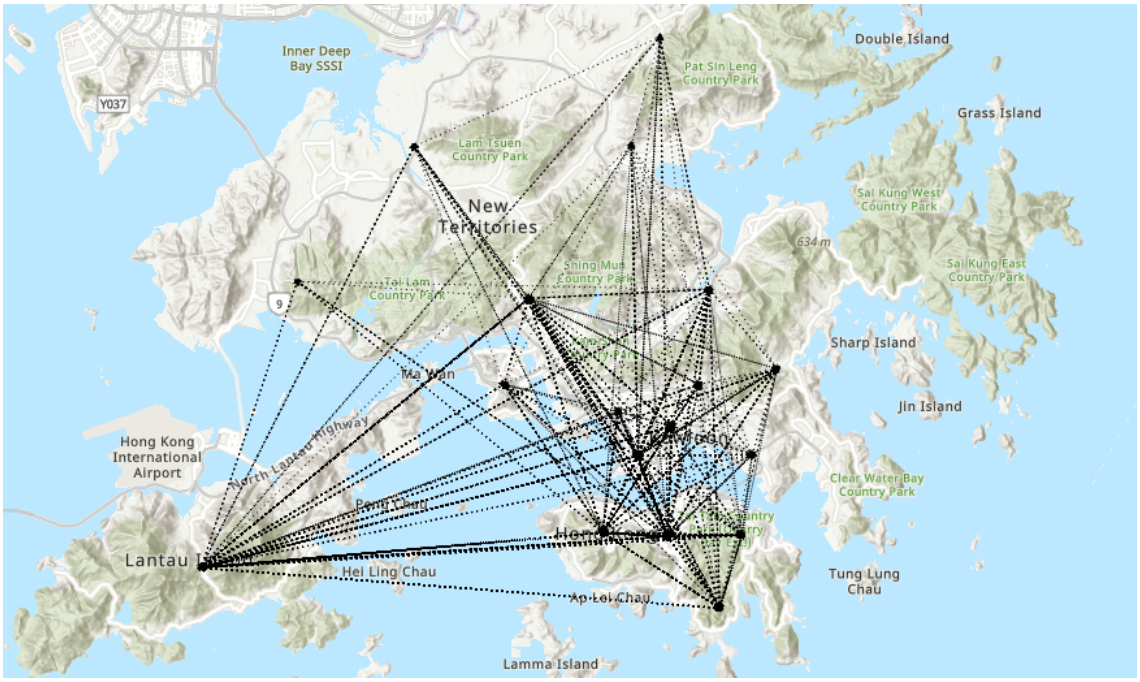


Figure 27. Example of the “Travel Network” layer.

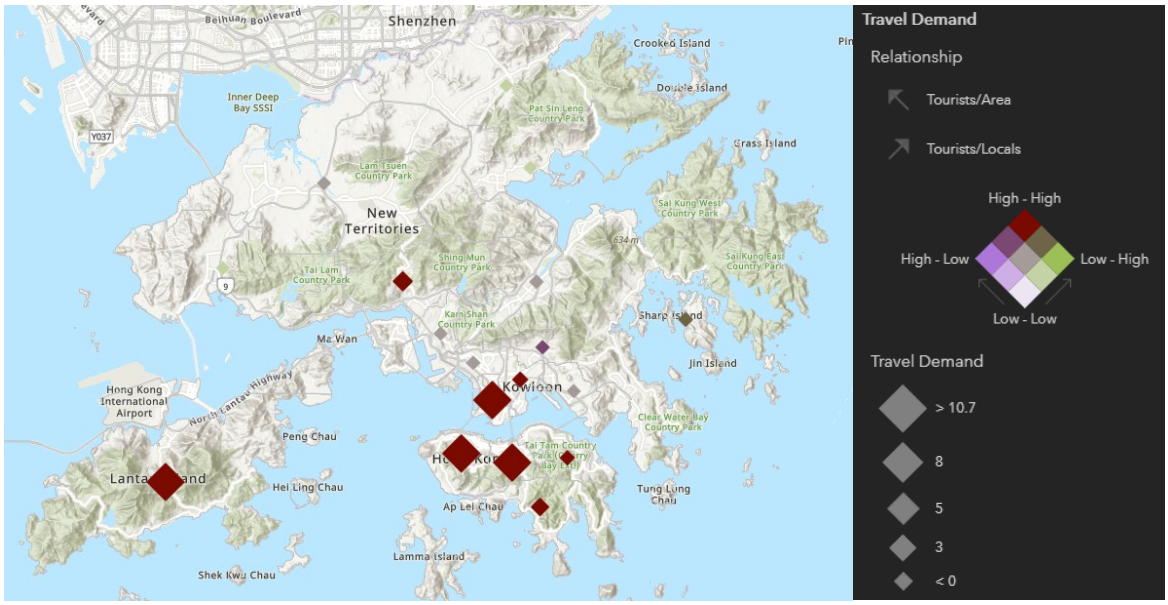


Figure 28. Example of the "Travel Demand" layer.

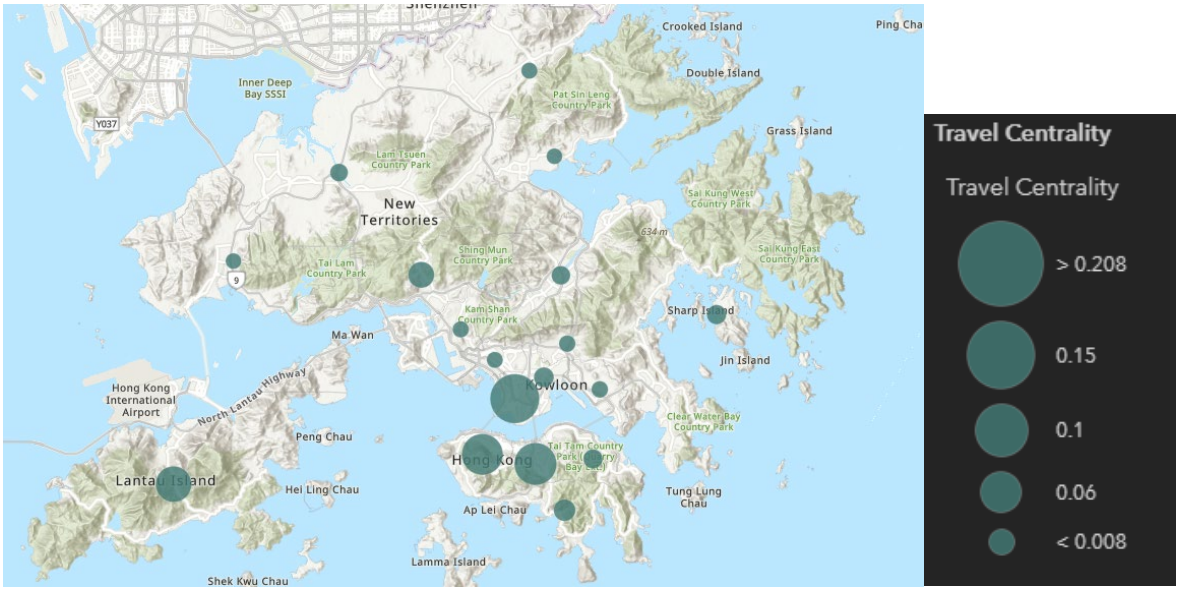


Figure 29. Example of the "Travel Centrality" layer.

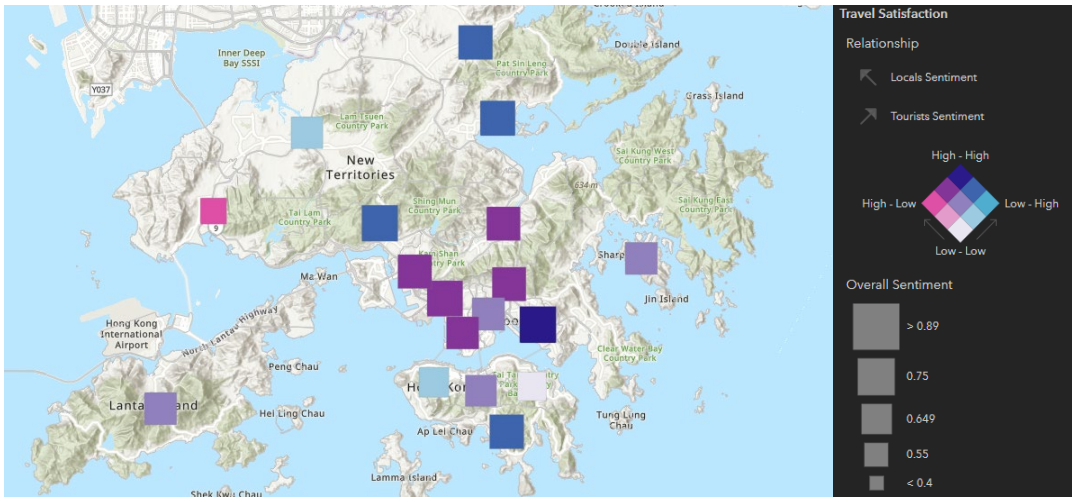


Figure 30. Example of the "Travel Satisfaction" layer.

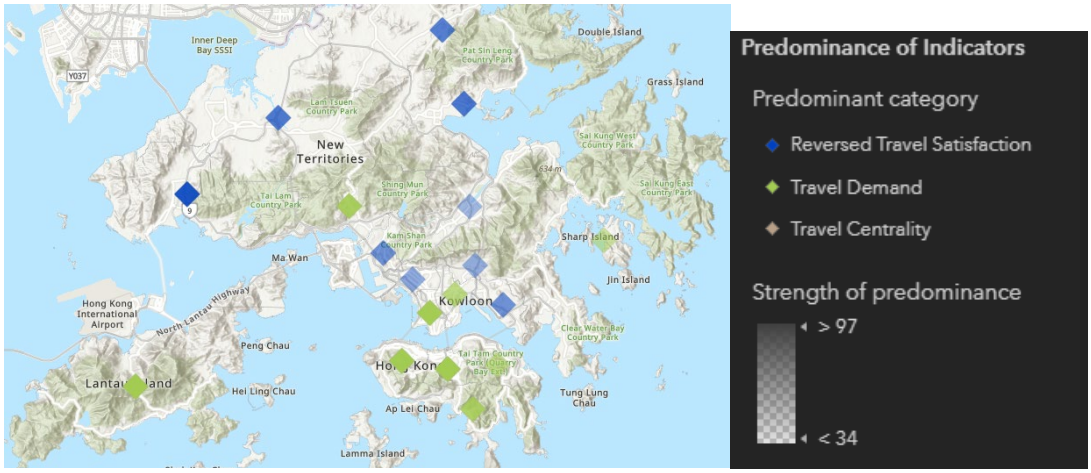


Figure 31. Example of the "Predominance of Indicators" layer.

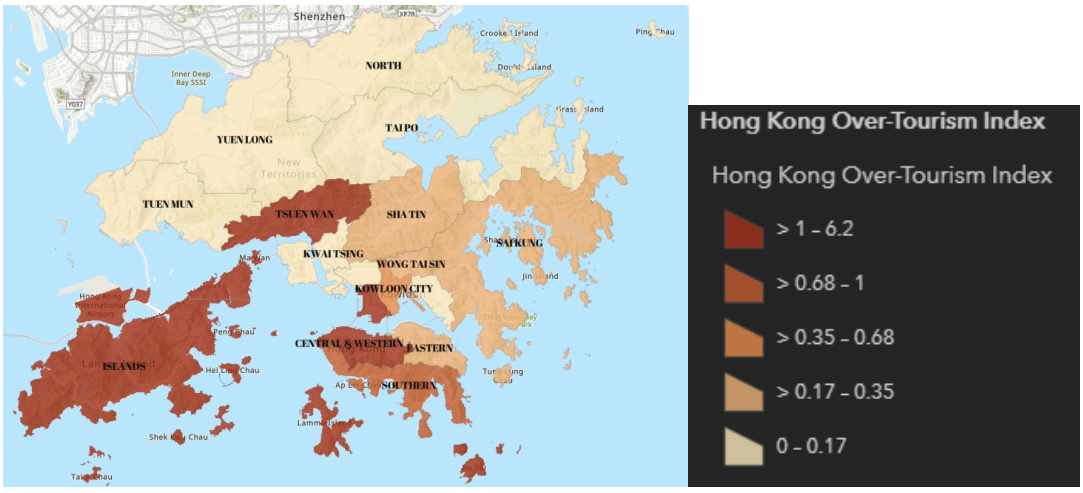


Figure 32 Example of the "Hong Kong Over-Tourism Index" layer.

5 Policy Implications and Recommendations

In the following part, the policy implications and recommendations are mainly based on the travel suggestions by the Hong Kong Tourism Board (HKTB)[1].

5.1 Reliving the Pressure of the High Travel Demand District

From the travel demand (TD) results in section 4.2.2, Islands, Yau Tsim Mong, Central & Western, and Wan Chai are the high TD districts; Tsuen Wan and Southern are in the middle; and other districts have similar minimum values. To relive the pressure of these high TD places, the best way is to transfer the tourists to other districts.

In the HKTB's "10 things every visitor must experience in Hong Kong[2]" suggestions, most of the activities or suggested places are concentrated in the high TD districts. Twelve out of fourteen places are in the above-mentioned four high TD districts. This is not the only case. Like clicking into picnic spots explore[3], eight out of nine places (Figure 33) are in Yau Tsim Mong and Central & Western. Another issue is that even though suggested spots are in the high TD district, the spots can be evenly distributed in the region or a little bit separated apart. But, from the map of the "local treasures of West Kowloon[4]" in Figure 34, fourteen stores are in the center of Yau Tsim Mong. This will cause visitors to congregate closely on these streets and create crowding problems. Therefore, it is necessary to revise the suggested route or locations of the Hong Kong Explore.

In the revision of Hong Kong Explore, it is better to spatially and timely suggest different locations to reduce the concentration of tourists. Considering the spatial problem, TD results, just

[1] https://www.discoverhongkong.com/eng/hktb/about.html
[2] https://www.discoverhongkong.com/eng/explore/iconic-hong-kong-experiences.html
[3] https://www.discoverhongkong.com/eng/explore/attractions/picnic-spots-in-hong-kong.html
[4] https://www.discoverhongkong.com/eng/explore/neighbourhoods/west-kowloon/made-by-hand-local-treasures-of-west-kowloon.html

like section 4.2.2, high, middle and low TD districts are presented. Hence, it can be acted as a reference to cut or increase the number of suggested spots in each district. Taking picnic spots as an illustration, other than Kowloon and Hong Kong Island, it can be in New Territories, such as Grass Island and Inspiration Lake. Another part is the temporal TD issue. Although no obvious pattern is observed presented in this study, it is believed that it does exist in Hong Kong. "Time" should be included in the consideration of planning recommendations. If a specific place always has a high demand of tourists in a period of time, the strength of suggesting that spot should be reduced. In short, high-demand tourists should be evenly recommended to the whole Hong Kong.

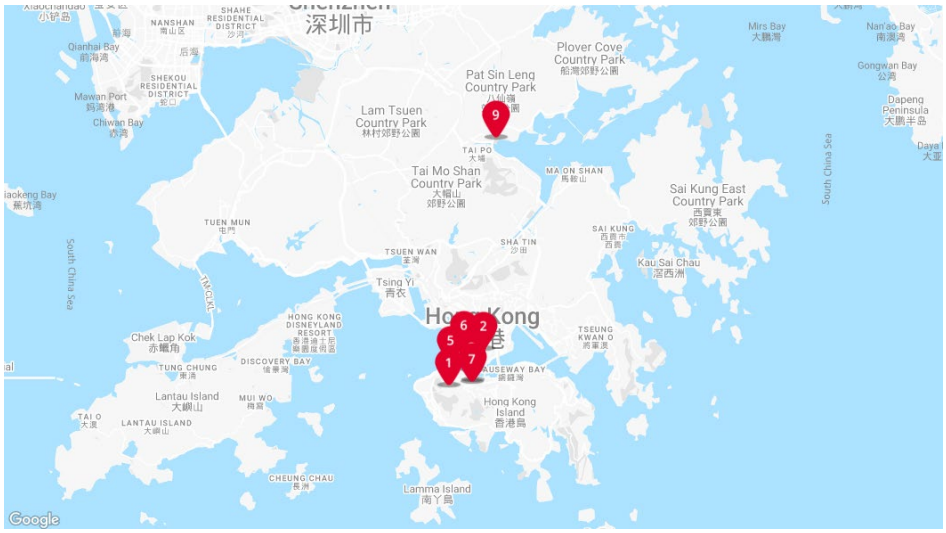


Figure 33. "Must-do" recommendation of picnic spots. (Source: Hong Kong Tourism Board (HKTB) (2022))

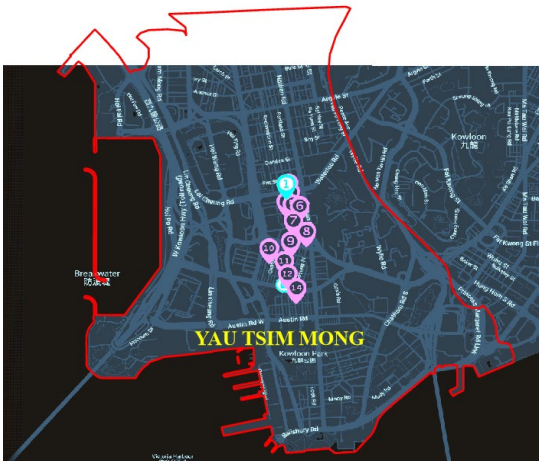


Figure 34. "Must-do" recommendation of local treasures of West Kowloon. (Source: HKTB (2022))

Additionally, a route/ trip planning system can be introduced in order to help the tourist crowding problem. Despite many pre-set suggested routes on the website, not all of them are suitable. The spot can be just a few people in yesterday but it becomes crowded today. Or there is an accident on the passing streets, causing congestion. Tourists should be reminded of the current situation of the place and their passing streets so that they can design their route. And the planning system should be able to notice the real-time events and thus suggest the best route. It could be not necessarily the shortest path, but is the most comfortable and suitable path. An examplee can be viewed in Figure 35, that blue and green lines are the recommended path regarding the level of crowding in different time periods. Various algorithms can help the development of this planning system (Cheng et al., 2013; Migliorini et al., 2021; Wu et al., 2017). Furthermore, the path can be adjusted according to tourists' interest. If the tourists would like to explore shopping malls, a path including nearby iconic malls can be generated. This not only can separate massive tourists, but also enhance their travel experience.

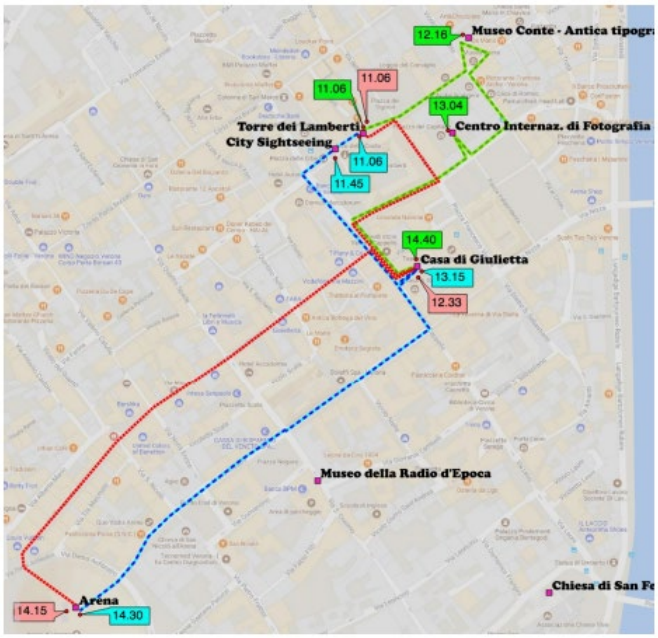


Figure 35. Different suggesting paths regarding the level of crowding. (Source: Migliorini et al. (2021))

5.2 Enhancing Tourists' and Locals' Satisfaction

In the TS results, section 4.2.3, both locals and tourists were having similar sentiment scores. They averagely felt positive. The factors that contribute to this feeling can be further investigated.

For each place, there must be a reason that causes a feeling. People use text to record it down. In this research, all the texts are transferred to a number only, without any categorical classification. In fact, this can be further developed in researching what specific item or event or place lead to that emotion. Then, these points would be the direction for the enhancement of tourists' and locals' satisfaction. For example, if most of the tourists reported negative feelings in Tsim Sha Tsui regarding the transportation, then revision on the transportation system may be needed. In contrast, if many visitors felt positive about Hong Kong's traditional food, then the promotion of Hong Kong's traditional food and food stalls can be emphasized. By matching the traveler's needs, the travel experience will become joyful. For locals, the analysis should be based on searching their range of tolerance and concerns. In terms of tolerance, it is known that residents would bow to certain benefits and accept the impacts brought by tourism. And the tolerance range and that specific event are localized issues so they should be individually studied for each community. Sports tourism is a good illustration. Bangsaen residents (in Thailand) are supportive to sports tourism (Boonsiritomachai & Phonthanukitithaworn, 2019) but Malaysians have the least interest in developing this type of travel (Chang et al., 2020). So, for Hong Kong, the development of suitable types of tourism is an important issue to be discussed, and the TS score together with the information behind it can provide clues to the discussion.

5.3 Revision on Transportation System and Tourism-related Facilities

In the TC results (section 4.2.4), the degree of the importance is computed for every district while travel network is plotted for indicating the from-to direction. Based on the this, transportation system and tourism-related facilities in each district should be reviewed. The rule of the revision should be according to the significance of the place. The more important the place, the more resources are invested. More vital point is to fulfill the needs of visitors and reduce the trouble to locals.

Transportation system is to satisfy the travel demand. It affects tourists' travel experience as people rely on it for traveling around. It also influences locals' living quality as it is the determinant of congestion issue, air-quality etc. The system planning should be aligned with needs. For instance, in the period of high OT sessions (2017/01 to 2019/07), Central & Western had a stronger connection with Yau Tsim Mong, Eastern, and Sha Tin. This implies that there was a high demand of travel flows and thus effective and efficient transport is necessarily needed to handle such high amount of people. And the travel network can help the planning of the transportation system in order to manage the massive tourists without influencing locals' daily lives.

Additionally, important places may need more support in dealing with visitors' issues Tourism-related facilities, including but not limited to accommodations, toilets, and information centers, place a vital role in tourists' experience, which directly affect tourists' satisfaction (Chen et al., 2013). So, these facilities should be both quantitively and qualitatively revised, especially for the iconic spots in high TC districts. Meanwhile, residents can also enjoy its advantages. The revision can lead to a win-win situation for Hong Kong.

5.4 Prioritizing the Tourism Strategies

Previously, reliving the pressure of the high travel demand districts, enhancing tourists' and locals' satisfaction, and revision on transportation system and tourism-related facilities are discussed with the focus on TD, TS and TC indicators. All three indexes are playing crucial role. But it is difficult to concentrate on all three at the same time for all districts. To prioritize the strategies, predominance of indicators is needed to reflect the dominant element in the OT index.

In the web-based decision support tool, a layer of "Predominance of Indicators" is used for this purpose. Spatial and temporal changes from 2017/01 to 2019/07 are plotted in Figure 37 while legend in Figure 36 helps the explanations of maps. It is obvious that those high OT districts (in dark red) were always largely influenced by TD while those comparatively low OT districts (in orange) were affected by TS. In addition, TC is the least impact element to Hong Kong. Based on this observation, corresponding policies can be studied and launched on the corresponding places so as to tackle the most urgent problem.

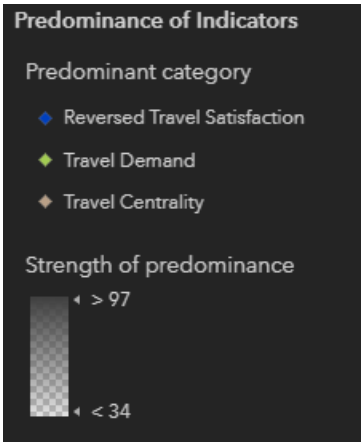


Figure 36. Legend for Predominance of Indicators.

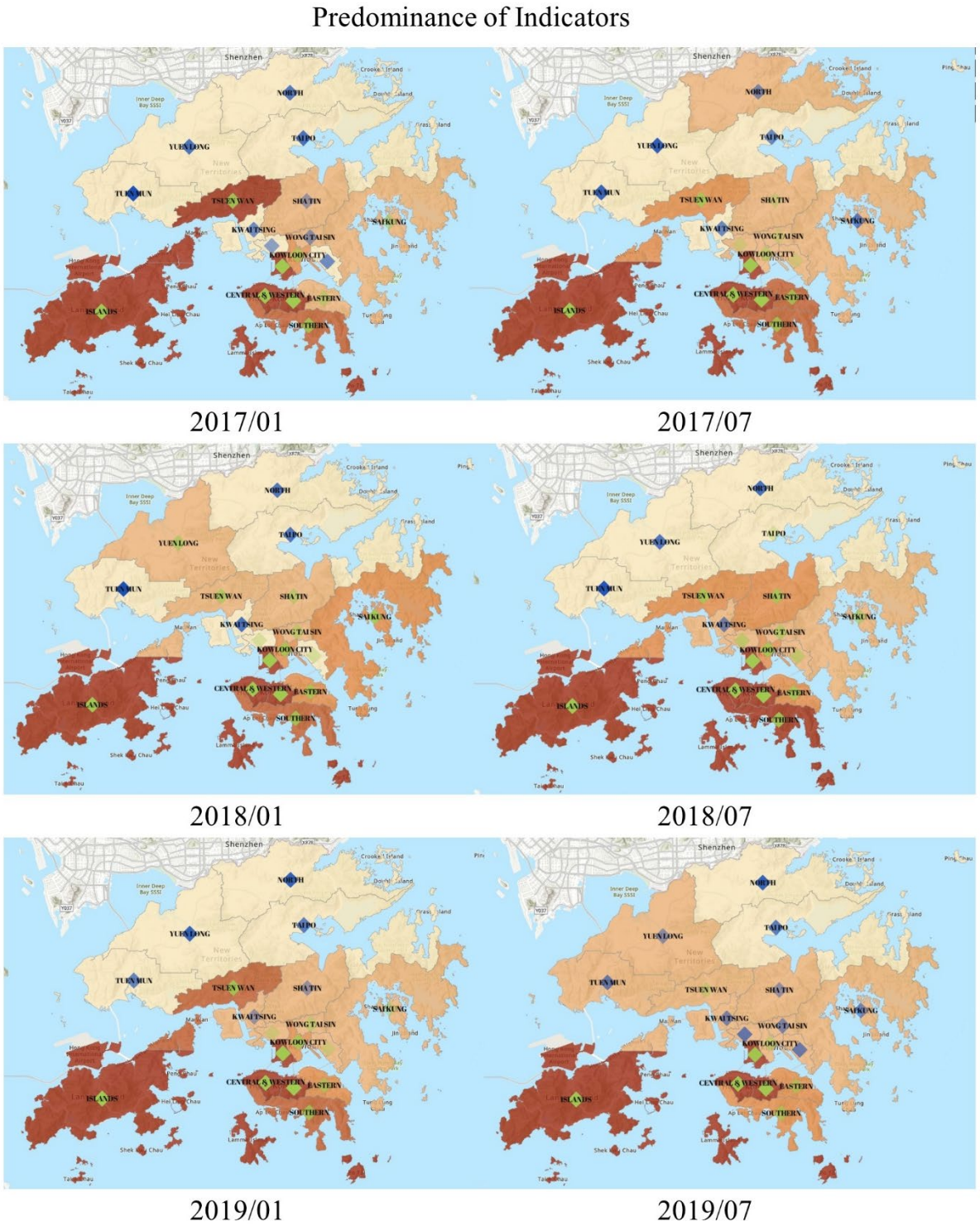


Figure 37. Predominance of Indicators.

5.5 Developing of a Tourism App

In this project, social media data is taken as the data source. Data acquisition is easier for academic purpose. However, if it is for governing, the data collection is much difficult or even impossible due to the user information protection. Just like in China, the forbidden of Weibo data to public is launched. So if the OT index would like to be modeled in the future, new data source has to be discovered. Currently, apart from social media data, another popular data is cellular data. Many research implements it in the tourism analysis, examples in (Chu & Chou, 2021; Wang et al., 2021). A defect of this kind of data is that sentiment analysis is not possibly to be performed due to the lack of textual content. To overcome this problem, a tourism app is suggested to be introduced.

For the app, it is advised to include the function of check-in with scoring. People are able to check-in the places and then share on different social media platforms. Moreover, tourists can rank a score to show their emotions towards the place and give feedback or comment. In this case, both sentiment and spatial-temporal information can be collected and further utilized for TD, TS and TC computation. In order to raise interest in using the app, special filters and stickers can be introduced for every special tourist hot spot. Actually, this can be a temptation to visitors to travel to designed places, like honeypot or less crowded locations. Separation of high concentration tourists could then be practiced.

Other than acting as a data source, this tool can greatly enhance tourists' travel experience. On the HKTB's website, there are many outdoor landscape routes (Figure 38). But the main problem is that it is on graphs rather than linking with the navigation system. This causes inconvenience to the user. If the route is linked with a navigation app, like Google Map, users can easily locate themselves and get the guidance of walking it. Furthermore, as mentioned in section 5.1, the route/ trip planning system can be one of the add-ins. Real-time and dynamic real-world situations detection can help the adjustment of the route and thus facilitate traveler's travel decisions. Tourists indeed also feel good about using the tourism app (Chang & Ahmad, 2022; Ramos-Soler et al., 2019).

Figure 38. Maps of outdoor landscape routes. (Source: HKTB (2022b))

To sum up, the app is just a transfer of the HKTB website, but with new few functions. And the advice of building a tourism app is to help the data collection and improve tourists' travel experience.

6 Public Dissemination

6.1 Online Public Platform

An online public platform, https://arcg.is/0mzHyH was developed to visualize the OT index and sub-indicators results. To enhance the user experience, this platform can be presented on desktop or mobile device. The public can access the platform without any limitations regarding devices, locations and time.

6.2 Research Paper

Based on the findings of this research project, a research paper is submitted to the Annals of Tourism Research. It is temporarily titled as “An Over-Tourism Platform Using Social Media Data: A Case Study in Hong Kong”. This paper aims at conveying the novel formation of OT index and introducing Hong Kong OT situation to the tourism academia.

6.3 International Conference

An international Conference, “The 12th Forum on Spatially Integrated Humanities and Social Science, Nov. 11-13, Guangzhou, China”, was held on 13 November 2022. The PI presented this research project in terms of the project scope, methodologies and research findings in the title of “Urban Sensing and Over-tourism in Hong Kong”. This conference was participated with academics in the aspects of geospatial, social science, computer science etc.

7 Conclusions

Over-tourism is a hot topic in all tourism cities as it is like a drug. It brings happiness to the city, such as economic growth and cultural output, but on the other hand, it destroys the sustainability of the city. To greatly gain the benefits and dispose of the bad, OT is a vital tourism issue to be tackled. In this research project, a thorough and comprehensive OT index model is set. Three indicators, travel demand, travel satisfaction and travel centrality, are developed in order to quantitively evaluate the social and environmental impacts. Based on the OT index model, the level of OT can be calculated.

Moreover, this research uses a novel approach for Hong Kong tourism issue evaluation. High-frequency data, social media data, is used. Unlike previous research using surveys or interviews as data sources, social media data can overcome problems of low-frequent, vague location and labor demand. The performance of the data meets the expectation. It can reflect Hong Kong OT situation at a certain extent. However, the restriction on data acquisition is getting strict. The forbidden of gaining Weibo data makes the research cannot perfectly present the estimation of actual Chinese tourists' arrivals. Meanwhile, collected Twitter data is only part of the Hong Kong Twitter posts because only non-precise geo-tagged posts can be obtained, which is only 1-2% of tweets (Twitter, 2022). So there is a discrepancy between estimated tourist arrivals and the actual tourist arrivals. This led to the fact that this research cannot perfectly model Hong Kong OT. In the future, if Hong Kong tourism app is developed, these problems can be resolved.

In future research, new data sources can be tried for OT modeling; three indicators, travel demand, travel satisfaction and travel centrality, can be deeply developed by adding more sub-components or using more advanced algorithms; the performance of the implementation of the OT index can be evaluated.

References

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). *XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond* (arXiv:2104.12250). arXiv. https://doi.org/10.48550/arXiv.2104.12250

Bianchi, F., Nozza, D., & Hovy, D. (2022). XLM-EMO: Multilingual Emotion Prediction in Social Media Text. *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 195–203.

Boonsiritomachai, W., & Phonthanukitithaworn, C. (2019). Residents' Support for Sports Events Tourism Development in Beach City: The Role of Community's Participation and Tourism Impacts. *SAGE Open*, *9*(2),

Caneday, L., & Zeiger, J. (1991). The Social, Economic, and Environmental Costs of Tourism to a Gaming Community as Perceived By Its Residents. *Journal of Travel Research*, *30*(2), 45–49.

Census and Statistics Department. (2019a). *Gross Domestic Product (Yearly) (2018 Edition)*. https://www.censtatd.gov.hk/en/data/stat_report/product/B1030002/att/B10300022018AN18E0100.pdf.

Census and Statistics Department. (2019b). *Year-end population for 2018*. https://www.info.gov.hk/gia/general/201902/19/P2019021900371.htm?fontSize=1

Census and Statistics Department. (2022). *Population by District Council District and Year (2021)*. https://www.census2021.gov.hk/en/main_tables.html

Chang, M.-X., Choong, Y.-O., & Ng, L.-P. (2020). Local residents' support for sport tourism development: The moderating effect of tourism dependency. *Journal of Sport & Tourism*, *24*(3), 215–234.

Chang, R., & Ahmad, M. Z. (2022). Identifying Service Items in Travel Mobile Applications Suitable for Senior Tourists to Improve Traveling Experience. *Mathematical Statistician and Engineering Applications*, *71*(3), 720–734.

Chen, Y., Zhang, H., & Qiu, L. (2013). *Review on tourist satisfaction of tourism destinations*. *3*, 13.

Cheng, S.-T., Chen, Y.-J., Horng, G.-J., & Wang, C.-H. (2013). Using Cellular Automata to Reduce Congestion for Tourist Navigation Systems in Mobile Environments. *Wireless Personal Communications*, *73*(3), 441–461.

Cheung, K. S., & Li, L.-H. (2019). Understanding visitor–resident relations in overtourism: Developing resilience for sustainable tourism. *Journal of Sustainable Tourism*, *27*(8), 1197–1216.

Chiu, H. Y., Chan, C.-S., & Marafa, L. M. (2016). Local perception and preferences in nature tourism in Hong Kong. *Tourism Management Perspectives*, *20*, 87–97.

Chu, C.-P., & Chou, Y.-H. (2021). Using cellular data to analyze the tourists' trajectories for tourism destination attributes: A case study in Hualien, Taiwan. *Journal of Transport Geography*, *96*, 103178.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, *80*, 56–71.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). Social Fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 1–1.

Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2017). Creating value from Social Big Data: Implications for Smart Tourism Destinations. *Information Processing & Management*, *54*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805*.

Dutheil, F., Baker, J. S., & Navel, V. (2020). COVID-19 as a factor influencing air pollution? *Environmental Pollution (Barking, Essex : 1987)*, *263*(Pt A), 114466.

Esri China (HK). (2021, April). *Hong Kong 18 Districts*. https://opendata.esrichina.hk/datasets/eea8ff2f12b145f7b33c4eef4f045513_0

Fang Bao, Y., & Mckercher, B. (2008). The Effect of Distance on Tourism in Hong Kong: A Comparison of Short Haul and Long Haul Visitors. *Asia Pacific Journal of Tourism Research*, *13*(2), 101–111.

Forster, J. (1964). The Sociological Consequences of Tourism. *International Journal of Comparative Sociology*, *5*(2), 217–227.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

Heung, V. C. S., & Quf, H. (2000). Hong Kong as a Travel Destination: An Analysis of Japanese Tourists' Satisfaction Levels, and the Likelihood of Them Recommending Hong Kong to Others. *Journal of Travel & Tourism Marketing*, *9*(1–2), 57–80.

Hiemstra, S., & Wong, K. K. F. (2002). Factors Affecting Demand for Tourism in Hong Kong. *Journal of Travel & Tourism Marketing*, *13*(1–2), 41–60.

Hong Kong Tourism Board. (2019). *Hong Kong Tourism Board Annual Report 2018/19*. https://www.discoverhongkong.com/eng/about-hktb/annual-report/annual-report-20182019/frontend/docs/HKTB_Annual_Report_eng.pdf.

Hong Kong Tourism Board (HKTB). (2022a). *Best family-friendly picnic spots in Hong Kong | Hong Kong Tourism Board*. Dhk-Local-Market. https://www.discoverhongkong.com/hk-eng/explore/attractions/picnic-spots-in-hong-kong.html

Hong Kong Tourism Board (HKTB). (2022b). *Great outdoors | Hong Kong Tourism Board*. Discover Hong Kong. https://www.discoverhongkong.com/eng/explore/great-outdoor.html

Immigration Department. (2022). *Total Visitor Arrivals*. https://partnernet.hktb.com/en/research_statistics/tourism_statistics_database/index.html

Jin, J. C. (2011). The Effects of Tourism on Economic Growth in Hong Kong. *Cornell Hospitality Quarterly*, *52*(3), 333–340.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323.

Liu, Q., Wang, Z., & Ye, X. (2018). Comparing mobility patterns between residents and visitors using geo-tagged social media data. *Transactions in GIS*, *22*(6), 1372–1389.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv Preprint ArXiv:1907.11692*.

Lozano, S., & Gutiérrez, E. (2018). A complex network analysis of global tourism flows. *International Journal of Tourism Research*, *20*(5), 588–604.

Ma, Q., Yuan, C., Zhou, W., & Hu, S. (2021). Label-Specific Dual Graph Neural Network for Multi-Label Text Classification. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3855–3864.

Mccool, S. (1994). Planning For Sustainable Nature Dependent Tourism Development. *Tourism Recreation Research*, *19*, 51–55.

McKercher, B., Ho, P. S. Y., & du Cros, H. (2005). Relationship between tourism and cultural heritage management: Evidence from Hong Kong. *Tourism Management*, *26*(4), 539–548.

Migliorini, S., Carra, D., & Belussi, A. (2021). Distributing Tourists among POIs with an Adaptive Trip Recommendation System. *IEEE Transactions on Emerging Topics in Computing*, *9*(4), 1765–1779.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* (Technical Report No. 1999–66). Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Pan, B., Zheng, Y., Wilkie, D., & Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Phi, G. T. (2020). Framing overtourism: A critical news media analysis. *Current Issues in Tourism*, *23*(17), 2093–2097.

PiuChan, M., Chan, C. W., & Kaale, J. (2018). *Economic and socio-cultural impacts of Mainland Chinese tourists on Hong Kong residents*. https://doi.org/10.1016/j.kjss.2017.11.004

Ramos-Soler, I., Martínez-Sala, A.-M., & Campillo-Alhama, C. (2019). ICT and the Sustainability of World Heritage Sites. Analysis of Senior Citizens' Use of Tourism Apps. *Sustainability*, *11*(11).

Russo, A., & Scarnato, A. (2017). "Barcelona in common": A new urban regime for the 21st-century tourist city? *Journal of Urban Affairs*, *40*, 1–20.

Shen, H., Luo, J. M., & Zhao, A. (2016). The Sustainable Tourism Development in Hong Kong: An Analysis of Hong Kong Residents' Attitude Towards Mainland Chinese Tourist. *Journal of Quality Assurance in Hospitality & Tourism*, *18*, 1–24.

Silva, B. N., Khan, M., & Han, K. (2020). Integration of Big Data analytics embedded smart city architecture with RESTful web of things for efficient service provision and energy management. *Future Generation Computer Systems*, *107*, 975–987.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Song, H., Li, G., Veen, R., & Chen, J. (2011). Assessing Mainland Chinese Tourists' Satisfaction with Hong Kong Using Tourist Satisfaction Index. *International Journal of Tourism Research*, *13*, 82–96.

Song, H., Veen, R., Li, G., & Chen, J. (2012). The Hong Kong tourist satisfaction index. *Annals of Tourism Research - ANN TOURISM RES*, *39*.

Song, H., Wong, K. K. F., & Chon, K. K. S. (2003). Modelling and forecasting the demand for Hong Kong tourism. *International Journal of Hospitality Management*, *22*(4), 435–451.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). *Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts*. 11.

Twitter. (2022). *Advanced filtering for geo data*. https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

Wang, A. H. (2010). Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. In S. Foresti & S. Jajodia (Eds.), *Data and Applications Security and Privacy XXIV* (pp. 335–342). Springer Berlin Heidelberg.

Wang, L., Wu, X., & He, Y. (2021). Nanjing's Intracity Tourism Flow Network Using Cellular Signaling Data: A Comparative Analysis of Residents and Non-Local Tourists. *ISPRS International Journal of Geo-Information*, *10*(10), 674.

Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49–72.

World Tourism Organization (UNWTO). (2018). *'Overtourism'? – Understanding and Managing Urban Tourism Growth beyond Perceptions*.

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *PLoS ONE*, *9*(5), e97010.

Wu, X., Guan, H., & Zhao, L. (2017). Tour route planning problem with consideration of the attraction congestion. *Acta Technica*, 10.

Xu, F., Nash, N., & Whitmarsh, L. (2019). Big data or small data? A methodological review of sustainable tourism. *Journal of Sustainable Tourism*, *28*. https://doi.org/10.1080/09669582.2019.1631318

Zhang, X., & LeCun, Y. (2017). Which encoding is the best for text classification in chinese, english, japanese and korean? *ArXiv Preprint ArXiv:1708.02657*.

Zhang, Y., & Wallace, B. (2016). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification* (arXiv:1510.03820). arXiv. http://arxiv.org/abs/1510.03820

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., & Du, X. (2019). UER: An Open-Source Toolkit for Pre-training Models. *EMNLP-IJCNLP 2019*, 241.