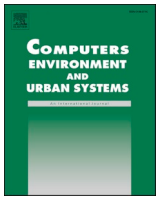
Computers, Environment and Urban Systems 85 (2021) 101567

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Spatiotemporal distribution of human trafficking in China and predicting the locations of missing persons

Yao Yao [a,c,1], Yifei Liu [b,1], Qingfeng Guan [a,*], Ye Hong [d,**], Ruifan Wang [a], Ruoyu Wang [e], Xun Liang [a]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China
[b] School of Resource and Environmental Science, Wuhan University, Wuhan 430075, China
[c] Alibaba Group, Hangzhou 311121, China
[d] Institute of Cartography and Geoinformation, ETH, Zurich, Zurich 8093, Switzerland
[e] Institute of Geography, School of GeoSciences, University of Edinburgh, Edinburgh, UK

ARTICLE INFO

Keywords:
Illegal adoption
Trafficking information network
Location prediction
Random forest
Public safety

ABSTRACT

In China, the illegal adoption of missing persons and especially of missing children is a major public safety issue that affects social and family stability. Recent work has established a trafficking information network developed from a volunteer-managed database of missing persons that identifies and locates node cities and critical paths of illegal adoption. In order to evaluate locations where trafficking can be identified and provide direct advice for affected families, this study analyses the temporal and spatial distribution of the missing population and explores factors that affect their transfer. We use spatiotemporal information to construct multiple random forest (RF) models for predicting the locations of missing persons transfer on a larger spatial scale. The proposed independent RF models, namely, provinces potentially entered, destination grids, relative distances and relative directions models, achieve high levels of accuracy. Moreover, an integrated RF-based city-level prediction model can effectively locate the city a missing person was trafficked to. From our driving factor analysis, the transfer paths are strongly correlated with source provinces and grids. The study also shows that the transfer of missing persons is driven by multiple factors rather than by a single element.

1. Introduction

The disappearance and trafficking of persons is a severe global problem (Rudolph & Schneider, 2014). In China alone, there are tens of thousands of disappearances every year (Wang et al., 2018). Most disappearances and transfers are essentially human trafficking (Qiu & Ma, 2015). Many scholars have been dedicated to studying the missing population and their transfer paths, mainly focusing on cross-border transfer, policies, legal developments, and law enforcement issues (Obokata, 2005). For example, Laczko and Ma (2003) described a series of European Union policies focused on missing populations and their transfer and posed several recommendations. In examining the case of missing persons in Britain, Kara (2011) proposed a more effective legal design framework for combating illegal trafficking. Brewster, Ingle, and Rankin (2014) proposed a new way of studying missing persons that involves mining information from open-source data (e.g., social media data) to better understand the experiences of this specific group of people.

The essence of most missing person transfer is human trafficking, which is a serious crime (Obokata, 2005). According to a study by Fajnzylber, Lederman, and Loayza (2002), income disparities, unemployment rates, and poverty levels have a more significant impact on the occurrence of criminal behavior. Moreover, Lochner and Moretti (2004) show that education has a strong impact on crime rates. Significant differences in economic, cultural, and educational levels can be observed across regions of China (Liu, 2014; Xu et al., 2005). Compared to other countries, China applies distinct administrative divisions and family planning policies, that is, the policy that one parent can only give

* Correspondence to: Q. Guan, School of Geography and Information Engineering, China University of Geosciences, 68 Jincheng Rd., Wuhan 320078, China.
** Correspondence to: Y. Hong, Institute of Cartography and Geoinformation, ETH Zurich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland.
E-mail addresses: yaoy@cug.edu.cn (Y. Yao), starry1004@whu.edu.cn (Y. Liu), guanqf@cug.edu.cn (Q. Guan), hongy@student.ethz.ch (Y. Hong), rickyse@cug.edu.cn (R. Wang), R.Wang-54@sms.ed.ac.uk (R. Wang), liangxun@cug.edu.cn (X. Liang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compenvurbsys.2020.101567
Received 21 June 2020; Received in revised form 8 November 2020; Accepted 9 November 2020
Available online 16 November 2020

birth to one newborn (Liu, 2003; Ma & Sun, 2011). Therefore, issues related to the missing population in China are unique (e.g., illegal adoption problems created by various regional birth policies) (Tian, 2012).

At present, the Chinese government has not publicly disclosed data concerning the missing population, and relatively few studies have focused on illegal adoption. Early studies such as Shen, Antonopoulos, and Papanicolaou (2013) have analysed the characteristics of the disappearance and transfer of the domestic population based on media reports, indicating that these activities are highly organized and that poverty is the main factor affecting them. Li et al. (2017) conducted a more in-depth exploration of the characteristics of population disappearance and transfer; at the spatial level, it was found that the majority of missing persons cases have occurred in the western regions and that missing persons are typically moved to the eastern region; at the temporal scale, the time periods in which missing persons cases have more frequently occurred was obtained. Nevertheless, the datasets used for these studies are small, and their time and spatial scales are rather rough.

The latest development has been Wang et al.'s (2018) spatial analysis of the missing population in China and on the establishment of a transfer information network. The work describes the overall spatial distribution in detail and identifies principal cities and pathways of population transfer (Wang et al., 2018). However, the study does not discuss or analyse the case of individual missing persons. The aforementioned studies have explained the causes, the distribution and transfer paths of missing persons, generating valuable results for policy formulation. Nevertheless, to date, no research has provided a means to predict the locations of missing persons.

This study involved an in-depth analysis of the spatial and temporal distribution of China's missing population. Baby Coming Back Home (www.baobeihuijia.com), established by non-governmental volunteers, is a large-scale information publishing platform for missing persons to find their homes and families to find missing persons. The data provided by the volunteers has high credibility (Wang et al., 2018). With data provided by Baby Coming Back Home and the random forest (RF) algorithm, we construct a destination prediction model to locate where missing persons may be located. First, a dataset for the missing population was constructed from self-reported data taken from the website. Based on this dataset, missing and transfer cases were analysed at different spatial and temporal scales. The major factors affecting the destination of the missing persons were then fed into RF-based classifiers, constructing models for predicting the whereabouts of the missing population. To generate more fine-scaled prediction results, we used statistical indicators to construct an integrated prediction algorithm allowing to predict potential missing person locations at the city level. The predicted results can be used to narrow down the search range in the searching process of the missing person's parents, public security agencies, and NGOs. We also explored the driving factors that affect the transfer locations of missing persons.

2. Methodology

2.1. Data preprocessing

We used public records provided on the "Baby Coming Back Home" website. The website is managed in cooperation with public security agencies, and Wang et al. (2018) has used the data source to analyse child-trafficking networks of illegal adoption in China. Therefore, we believe the data to be valid and highly reliable. Through a web crawler application, we acquire 35,434 raw records from the website. These data are derived from the self-reported disappearance process of the missing person several years after the event, which contains detailed information about the disappearance scene, to find his or her original family. Each data record includes information of the missing person, including the date of birth, gender, height, year and month of disappearance and missing and arriving locations. After excluding records with missing or erroneous data (such as the conflict in birth and missing date or unreasonable height), a total of 24,833 valid records were found.

2.2. Implementation of the prediction model

We then predict the locations of missing persons and analyse what are the driving factors of trafficking. The following steps were used for our predictions (Fig. 1): 1) Raw data is pre-processed and spatial-temporal features are constructed; 2) Independent position prediction models are built using these features and multiple rough ranges that the missing person may be transferred to are obtained; 3) a decision-making process is proposed to construct an integrated prediction model, where the probability of the missing person being located in certain cities is determined, and accuracy evaluations are carried out.

Our goal is to output a more precise geographic range for identifying missing persons. However, due to the limited number of training samples available and the large number of prefecture-level administrative districts in China, directly training urban models can lead to a reduction in generalizability (Zhang, Ballas, & Pineau, 2018). Considering the small differences in customs and cultural differences on the actual city boundaries, the "city" defined entirely following the division of administrative regions has apparent limitations. Direct city-level prediction results are more prone to over-fitting, and larger discrete errors will likely to occur. Hence, predictions were first made on a broader scale, and then each model is combined to obtain a more accurate range.

2.3. RF-based independent prediction models

In this section we present multiple independent prediction models and predict different aspects of arrival locations to enhance generalization.

We build a model to predict provincial administrative regions where missing persons may arrive and refer to this model as the "provincial prediction model." Arrival and disappearance locations are spatially related, and a "relative distance model" and "relative direction model" are established and refer to as the distances and orientations of the arrival locations relative to places of disappearance. Also, we divide China into regular blocks and establish a "grid prediction model." In turn, four independent predictive models are constructed as shown in Table 1.

The location to which the missing person was transferred is related to the missing location, age, year and month, gender, and height of the missing person. This information related to the transfer of missing persons is defined as missing features, which can be divided into spatial features, temporal features, and auxiliary features. Specifically, the missing and arrival locations are spatial features; the year and month of birth and missing are temporal features; while gender and height belong to auxiliary features.

We first consider the examined provinces (including autonomous regions, municipalities and special administrative regions) in describing spatial characteristics. However, truncation errors in the border areas of provinces will greatly affect the downstream tasks. More specifically, in border areas belonging to different provinces, the environmental customs, dialects, and public security situations are similar. Therefore, the grid division model is proposed to revise and supplement the classification criteria of the provincial model. The definition and division rules of the grid model are described as follows:

We use the minimum and maximum longitude and latitude values of disappearance and arrival locations as the range of the surface quadrilateral and divide this into 256 grids (16 horizontal and vertical divisions). After removing the grids with no origin or destination data, a total of 107 active areas remain and are used in this study, as shown in Fig. 2. The number of cities in different grid locations is different. In the southeastern region of China, the number of cities in each grid can reach five or more, but in the western region, multiple grids may only cover

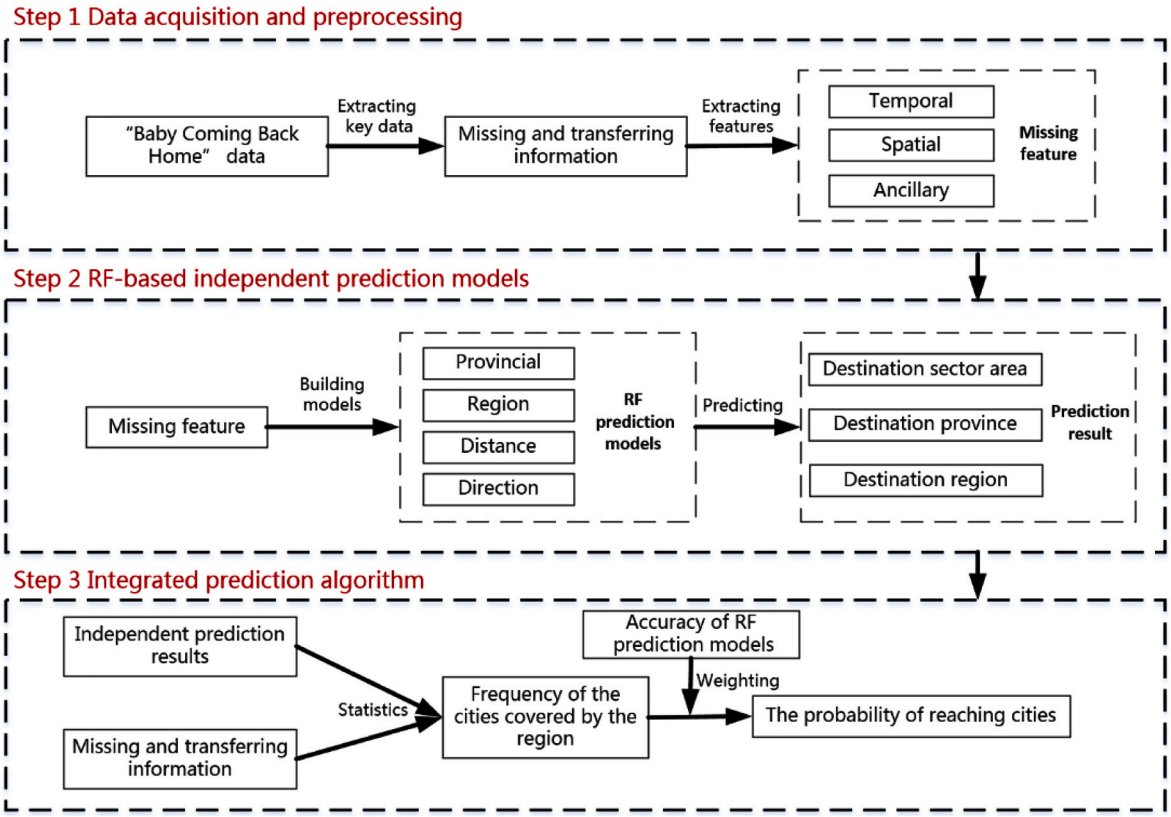


Fig. 1. Flowchart of integrated forecasting model implementation.

Table 1
The four independent prediction models used and their descriptions.

Model	Details
Provincial	The province where a missing person was trafficked to
Grid	The grid where a missing person was trafficked to
Distance	The reported distance from place of disappearance
Direction	The reported relative direction from place of disappearance

one city. On average, the number of cities per grid is 3.11.

Using the position of missing as the origin in a polar coordinate system, the arrival location can be described with relative distance and orientation. We used the Euclidean distance between disappearance and arrival locations as the transfer distance (polar diameter) and divided it into 20 categories using the Jenks natural break method (Chen et al., 2013). This classification method can limit differences within classes and ensure differences between classes. The relative transfer direction (polar angle) is used to determine the position interval of each transfer, which is divided into 8 direction includes north, northeast, east, southeast, south, southwest, west and northwest. Definition of transfer distance and relative orientation are shown in Fig. 3.

Temporal characteristics associated with disappearance include the time of disappearance and the age of the missing person. We use the year and month to describe the time of disappearance. At the 1980s, China began to implement the policy of reform and opening up. With the economic development, the crime rate increased significantly (Wang, 2015), so the illegal adoption of the people became rampant. Therefore, we only consider years following 1980. The remaining years are binned into five-year interval. Moreover, missing months are classified into Spring (March to May), Summer (June to August), Autumn (September to November) and Winter (December to February) seasons to account for seasonal effects on illegal trafficking.

The social abilities of different genders differ. Schoon and Polek (2011) note that the gender of a missing person needs to be taken into consideration. Moreover, the height of a missing person affects the occurrence of trafficking and the population transfer process because people with taller heights are more complicated to be transferred and transported. All the abovementioned features are used as input data for the proposed model.

The implementation of the independent prediction model relies on the use of the RF method. An RF model integrating multiple decision trees is used as a bagging tool in machine learning (Svetnik et al., 2003). The RF model is currently the best-performing high-dimensional nonlinear fitting model and can effectively avoid correlations between high-dimensional features (Fern, Ndez-Delgado, et al., 2014). The model has been widely used in the field of geography and in position prediction research and has proven useful for solving geographic problems with spatial complexity (Liu et al., 2017; Yao et al., 2018; Yao et al., 2019). When training the RF model, a commonly adopted procedure to increase the generalization ability is to randomly select 60% of the data as a training set at each time, that is, 14,899 out of 24,833 records are randomly selected to construct a decision tree. The remaining 40% of the data is used as a validation set to evaluate the generalization accuracy of the model. Following this procedure, 100 decision trees are built to form the an RF model. We construct four RF models, namely, provincial, grids, relative distance, and relative direction prediction models. With the same input features (i.e., the age, height, gender and year, month, province and grid of missing), the four prediction models output the province, grid, relative distance, and relative direction that the missing person transfers to, respectively.

2.4. Integrated prediction algorithm

According to each independent prediction model, the probability of the transfer provinces, grids, transfer distance and direction of a missing person can be obtained. However, the accuracy of each prediction model

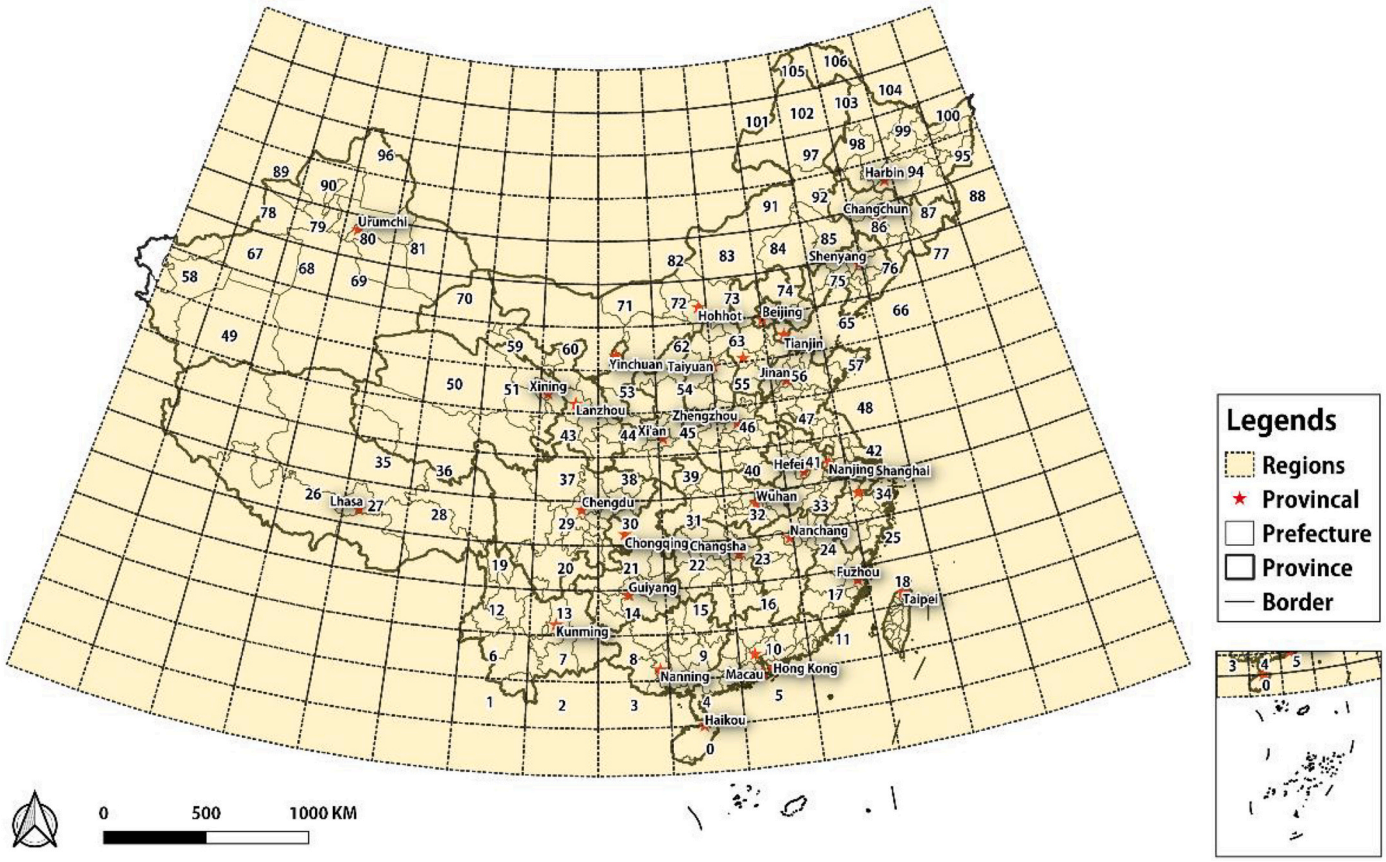


Fig. 2. Schematic diagram of grids partitioning excluding grids with no reported trafficking.

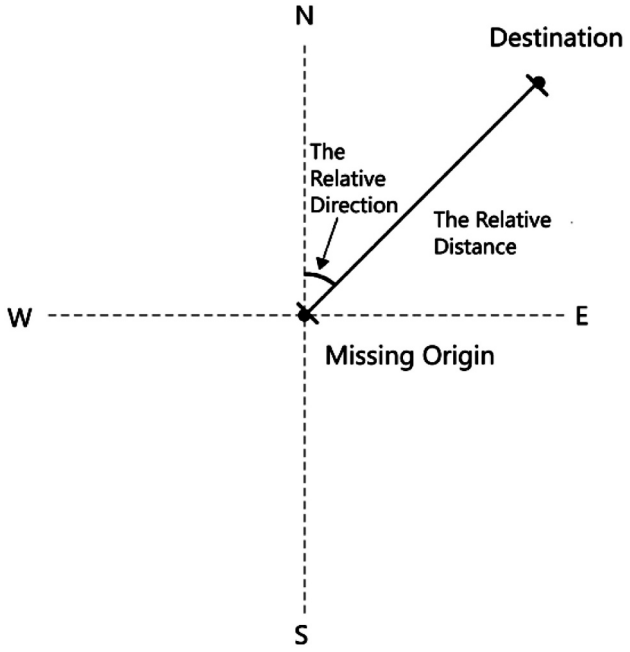


Fig. 3. Definition of distances and relative directions of trafficking destinations and origins.

is limited to make an accurate fine-grained prediction. We further propose a comprehensive prediction algorithm that combines the independent prediction models.

The results obtained by the province and grid prediction model are based on Cartesian coordinates. These predictions give a clear range of latitude and longitude coordinates where the missing person may arrive. However, the prediction results of the distance and direction prediction models are based on the polar coordinate system with the missing location as the origin. To further refine the relatively broad prediction ranges, we combine them into a sector area, which is shown in Fig. 4.

We first identify cities included in this sector area. When the area

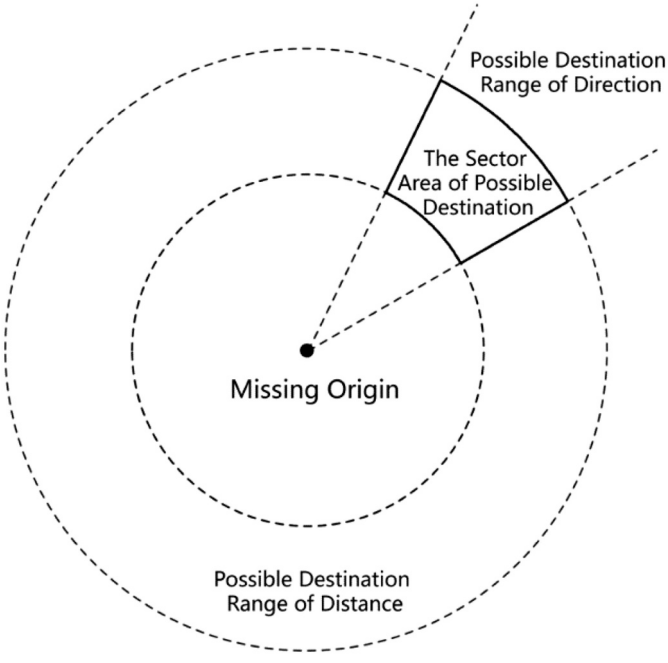


Fig. 4. Sector areas of possible destinations obtained by integrating the direction and distance range results.

covers n arriving cities included in the sample set, which are labeled as *city* 1, *city* 2, …, *city* n, and when the frequency of occurrence in the sample set is Q_{s1}, Q_{s2}, …, Q_{sn}, after arriving in this sector area, the occurrence probability F_{si} of city i is as follows:

$$F_{si} = \frac{Q_{si}}{\sum_{i=1}^{n} Q_{si}} \tag{1}$$

Similarly, the likelihood (F_{pi} and F_{ri}) of a missing person entering city i after reaching a province or grid can be obtained.

Based on the results of the predictive model, the probabilities of direction and distance ranges, provinces, and grids can be determined. The probability of reaching a specific sector is obtained from direction and distance probabilities. Since the direction probability P_{dire} and distance probability P_{dist} are independent, a sector's probability P_s is calculated as follows:

$$P_s = P_{dire} \cdot P_{dist} \tag{2}$$

The probability of reaching province P_p and the probability of reaching grid P_g can be directly obtained from the results of the independent prediction model.

According to the conditional probability model measuring independent events, the probability of reaching a city (P_{citys}, P_{cityp}, and P_{cityg}) can be calculated from the distance and direction, provincial, and grid models, respectively. We also use the overall accuracy (OA) to measure the model's predictive ability (Thapa & Murayama, 2009), which proved to be very effective in socio-economic, remote sensing and geography studies (Liu et al., 2017; Yao et al., 2018). OA values of the sector, provincial, and grid prediction models are $OA_s = OA_{dire} \cdot OA_{dist}$, OA_p, and OA_r, respectively. Taking the probability of reaching city i according to the sector as an example:

$$P_{city_{s_i}} = P_s \cdot F_{si} \cdot OA_s \tag{3}$$

The confusion matrix and hence the Kappa coefficient can be calculated with the test data for each prediction model. Kappa coefficient is a robust measure of models performance that balances error and accuracy. Kappa coefficient is often used in classification accuracy evaluation (Stehman, 1996). The value is in the range of 0 to 1. More than 0.5 indicates higher significance, and the closer the value is to 1, the higher the classification accuracy. We use the Kappa coefficient to weight the prediction results of the model, that is, the results predicted with higher Kappa score should have a higher weight in the comprehensive prediction algorithm. Thus, The probability of reaching a city as determined from the above formula is weighted according to the Kappa coefficient to obtain the final probability of arriving in a given city. Kappa coefficients for sectors, provinces, and grids are recorded as K_s, K_p, and K_g, respectively. Finally, the probability of reaching city P_{city} is as follows:

$$P_{city} = \frac{P_{city_s} \cdot K_s + P_{city_p} \cdot K_p + P_{city_g} \cdot K_g}{K_s + K_p + K_g} \tag{4}$$

P_{city} can be regarded as the final result of the integrated forecasting model.

3. Results

3.1. Spatiotemporal analysis of human trafficking

3.1.1. The features of missing persons at the yearly scale

Fig. 5 shows the number of people who are reported missing each year according to our dataset. The earliest cases of missing persons reported date back to 1926. More disappearance cases occurred in the first two stages of China's reform and opening period. From 1979 to 1992, the number of missing persons cases reached 13,021, accounting for 52.43% of the total number. According to statistical results of the

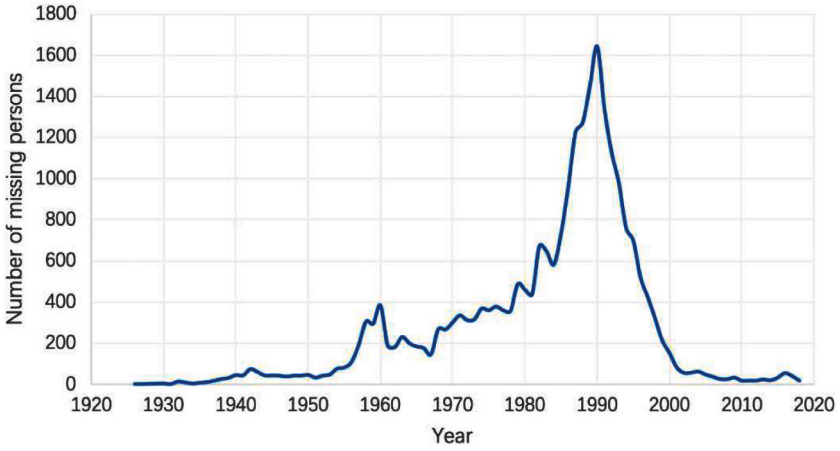


Fig. 5. Number of missing persons recorded for each year.

missing population dataset, missing children of 0–3 years of age accounted for 83.92% of the total missing population, and the proportion of children (0–6 years of age) reached 94.44% while the proportion of juveniles (14 years of age and younger) reached 98.98%. Adults (18 and older) account for only 0.51%. Therefore, the majority of missing persons are children.

These phenomena have mainly occurred for the following three reasons: From 1959 to 1961, China experienced a "three-year natural disaster." Low levels of social and economic security led to an increase in the number of crimes committed, which affected the number of missing persons. In the 1980s, at the start of the reform and opening policy period, disorder created by changes in China's social structure led to a significant increase in missing persons cases. Afterward, levels of social security gradually improved, and the number of missing persons cases decreased (Ni et al., 2008). The gradual and in-depth implementation of family planning policies has also greatly affected the occurrence of this

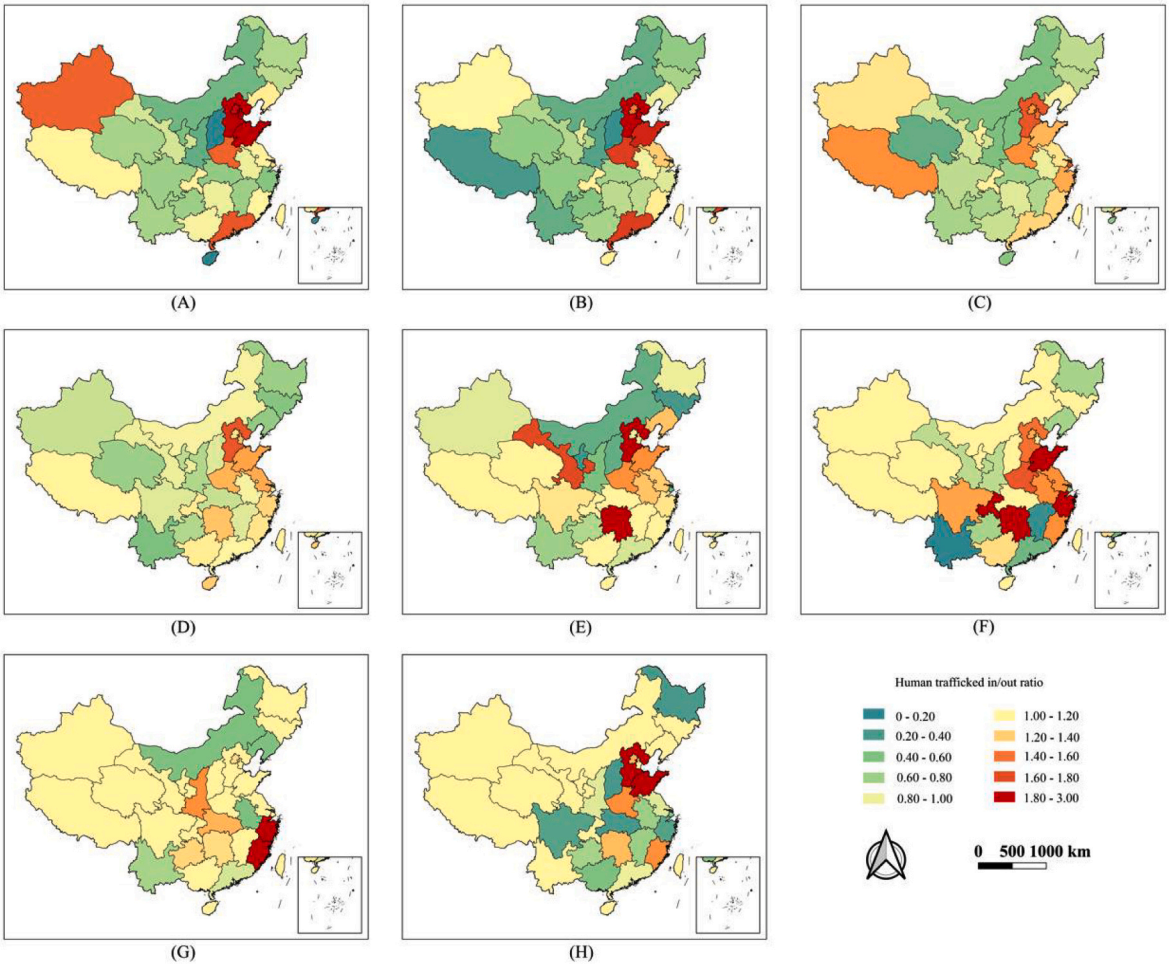


Fig. 6. Ratio of individuals human trafficked in and out of each province on a yearly scale: (A) 1980 to 1985, (B) 1985 to 1990, (C) 1990 to 1995, (D) 1995 to 2000, (E) 2000 to 2005, (F) 2005 to 2010, (G) 2010 to 2015, and (H) 2015 to 2018.

criminal act. Since their application in 1991, most of the newborns are the only child in the family, thus receive more attention from their parents (Ma & Sun, 2011), resulting in a decrease in the number of missing persons.

Levels of public security, transportation conditions, and national policies vary in different eras, which directly or indirectly affects the conditions of the missing population, thus generating different spatial distribution results. We focus on missing persons cases occurring after the year 1980 and apply five-year time segments. Fig. 6 shows the net inflow rate of missing persons (e.g., the ratio of individuals human trafficked in and out of each province in each five years division).

The inflow of missing persons into Hebei, Shandong, and Henan was significantly greater than outflows out of these provinces most of the time. Inflows and outflows to and from Hunan, Zhejiang, Jiangsu, Chongqing, and Gansu are relatively balanced. Most regions of northeastern and southwestern China show higher outflows than inflows.

Since 1979, the implementation of China's basic national policy (reform and opening up) has affected the distribution of missing persons. During this period, there was a significant directional change in the social structure of Guangdong province relative to other parts of the country (Chen, Zheng, & Jia, 2017), leading the outflow of missing persons to gradually exceed inflows.

3.1.2. *The features of missing persons on a monthly scale*

Monthly information on periods in which individuals are missing is extracted for each record included in the dataset. The climate, habits, and customs in different months will have a certain impact on the process of human trafficking and transfer. Fig. 7 shows the number of missing persons cases recorded for every month.

In terms of conditions observed in each year, conditions in January have been the worst, representing a total of 2766 missing persons cases. In particular, the number of missing persons cases recorded before the Chinese Lunar New Year (corresponding to January and February) and afterward (corresponding to March) exceeds that for the other months, as the flow of people trafficked during this period was relatively large (Liu & Shi, 2016), and an increase in the migrant population led to an increase in crime rates (Wang, 2016). As a result, the number of missing persons cases in these months is higher than that of other months. Second, missing persons cases are more common in October, June, May and April, when local climatic conditions are milder, representing 33.58% of all cases. The climate has an impact on the transfer missing persons, and the occurrence of extreme weather events occurrence in these months is low (Zheng, Wu, & Wang, 2014), which is more conducive to population transfer. In other months, the occurrence of extreme weather patterns and temperatures is not conducive to human trafficking.

Climatic conditions and cultural customs occurring in different months of the same year affect the transfer of the missing population and thus its spatial distribution. Fig. 8 shows the mean transition path lengths for missing persons cases for each month.

The average transfer path length fluctuates from 340 to 420 km. The maximum length is observed in January (412.38 km) while the minimum length is observed in July (344.62 km). The long-distance migration of residents before the Spring Festival has also affected the transfer path lengths of the missing persons (Zhao & Wang, 2017). Due to relatively mild climatic conditions occurring in April, May, June, and September, missing persons can be more easily trafficked in these months, lengthening the transfer paths in these months. By contrast, in February, March, July, and August, when extreme weather conditions are more frequent, more consideration of transportation costs and pathogenic factors is required (Noort et al., 2012).

3.2. *Predicting the destinations of missing persons*

After feature extraction and classification, we adopt RF classifiers to construct independent prediction models on arrival provinces and grids, relative distances, and relative orientations. The results of the proposed model are analysed, and factors that influence the whereabouts of missing persons are explored (Archer & Kimes, 2008).

In synthesizing the node growth and division of the decision tree in the RF, parameter weights included in the independent prediction models are shown in Table 2. Provincial, grid, distance, and relative direction models are respectively represented in the table and refer to provincial, grid, distance, and direction-related data.

Regardless of which model is used to predict the arrival locations of missing persons, dominant factors concern from where a person is reported missing, including the corresponding province and grid, followed by the age and the year and month of disappearance. The gender and

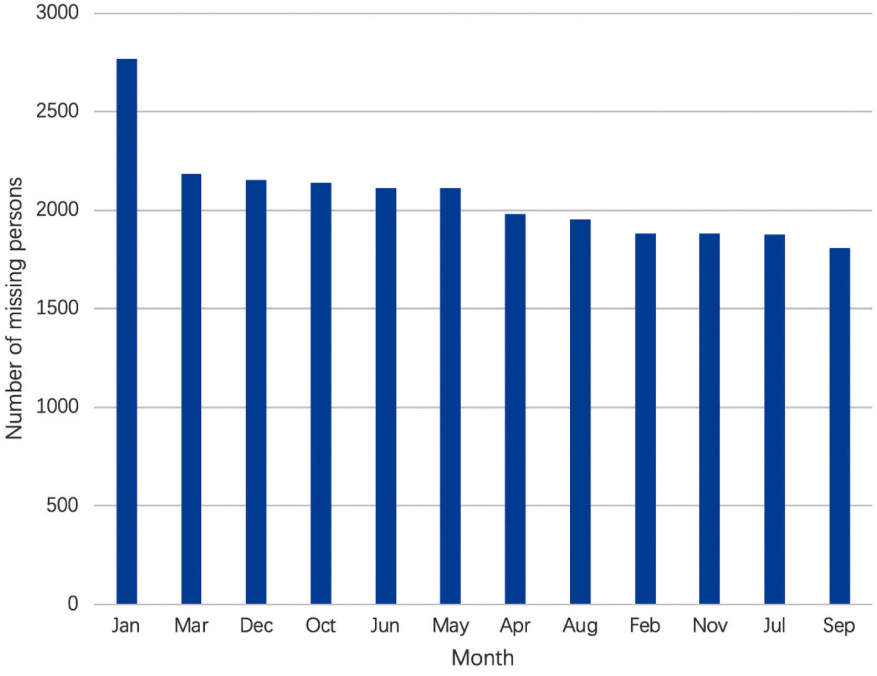


Fig. 7. The distribution of the total number of missing persons in each month from 1926 to 2018.

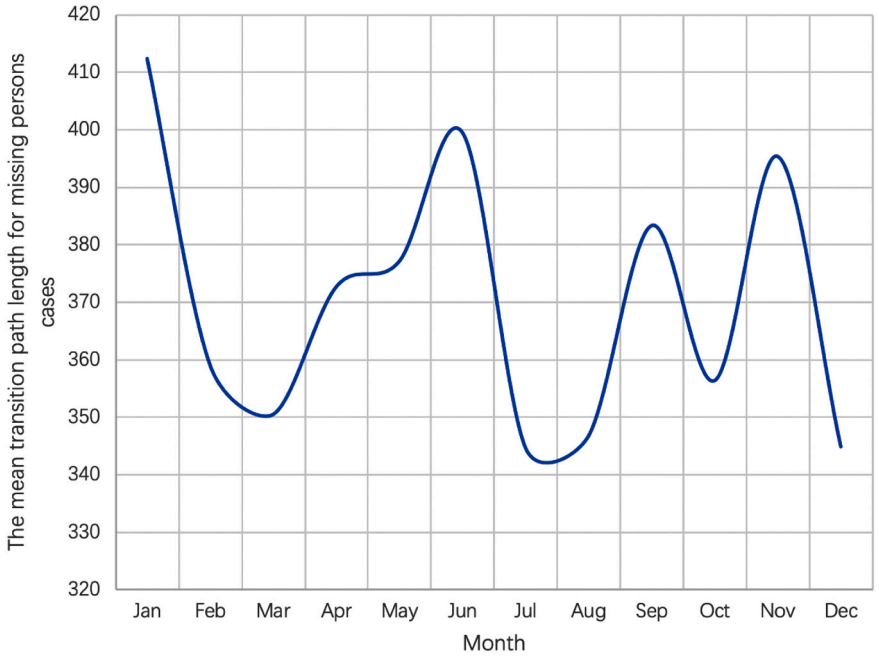


Fig. 8. Mean transition path lengths for missing persons cases for each month.

Table 2
The weights of various features included in each independent prediction model. (The values in the table that favor warm (red) background colors represent higher weights, and values that favor cool (green) background colors represent lower weights).

height of a missing person are auxiliary factors that have limited influence on the locations of transfer. The sum value of these two factors only accounts for roughly 20% in each model.

The prediction of destinations can only be verified by real data evaluations, and we are unable to create verification data or to manually discriminate model errors. As the amount of original data available is limited, we use the validation error to evaluate the performance of the proposed models.

The output of each model includes the predicted results and their corresponding probabilities. We regard the most likely predicted result as the predicted result of the models. In assessing the accuracy of each model, we used the overall accuracy (OA) value and Kappa coefficient as shown in Table 3.

As a result, the OA and Kappa coefficient values of the provincial and grid prediction models are relatively high (the Kappa coefficients of both models exceeding 0.7) while those of the relative distance and direction models are relatively low. Different policies implemented by different provinces and grids affect the transfer of missing persons (Wang et al., 2018), which is more comprehensively considered in the provincial and grid prediction models.

The results of the integrated city prediction model reveal potential destination cities and their corresponding probability values. For each prediction, the integrated model returns several potential arrival cities. These cities are sorted according to their corresponding probabilities from high to low.

We select known data from the training set to test and define the following test rules: 1. First Recall: the first city included in the result set corresponds to the actual result.; 2. Top 3 Recall: the correct results appear in the first three predicted cities; 3. Top 5 Recall: the correct results appear in the first five predicted; 4. Top 10 Recall: the correct results appear in the first ten predicted cities. The accuracy of the evaluation results of the integrated prediction model is shown in Table 4.

The results show that when applying the integrated prediction algorithm, missing persons have a 39.50% probability of appearing in the first city and a 82.99% probability of appearing in the top ten cities.

The above results are calculated using the comprehensive prediction model. While the city-level prediction model can only obtain 64.65% of the top 10 recall, the comprehensive prediction model can reach a top 10 recall of 82.99%, meaning the prediction accuracy has been greatly improved.

4. Discussion

Human trafficking is a critical issue that affects family harmony and social stability. In consideration of levels of social development, the implementation of national policies, and different climates and customs, we have innovatively analysed missing persons in China from a spatio-temporal perspective. Also, we use an integrated model to predict cities

Table 3
OA and Kappa coefficient results for each prediction model.

Prediction model	OA	Kappa
Provincial	0.800	0.786
Grid	0.757	0.745
Distance	0.639	0.581
Direction	0.656	0.608

Table 4
Accuracy of the evaluation criteria and results of the integrated prediction algorithm.

Assessment criteria	Rates
First Recall	39.50%
Top 3 Recall	63.56%
Top 5 Recall	73.02%
Top 10 Recall	82.99%

that missing persons may be found in.

This result shows that at the yearly scale, the conditions of missing persons vary due to varied national policies and social conditions. The number of missing persons cases has decreased significantly in recent years. At the monthly level, more disappearances have occurred around the Spring Festival, when large-scale patterns of migration occur across China. In periods with milder climatic conditions, the number of missing persons cases is higher, with longer path of transfer. Hence, the human trafficking patterns are closely related to temporal and spatial characteristics.

We propose a model for predicting the locations of missing persons and provide a precise and feasible method. The proposed model extracts spatial, temporal, and auxiliary features of missing cases and transform predictions of arrival locations into a multiclass prediction problem. In applying the first law of geography (Sun et al., 2012), to attenuate discrete errors resulting from the use of different location methods, four location division methods are used to describe where missing persons may be located.

This study innovatively uses latitude and longitude networks to divide the locations of missing periods and adopt the RF method, which offers strong generalization performance, to establish prediction models. The proposed models are comparatively accurate, with the OA and Kappa coefficients of the provincial model reaching values of 0.800 and 0.786, respectively. The OA and Kappa values of other predictive models exceed 0.639 and 0.581, respectively.

This study combines the probability results of each independent model with the frequency of arrival cities and weight them by accuracy to obtain a set of city-level prediction results. The accuracy of our city-level results is high, presenting a recall rate of 82.99%. Compared to the independent RF prediction model, the city-level integrated algorithm is more generalized and provides fine-scaled location information. Since very young missing persons (less than three years old at the time of missing) accounted for 83.92% of the original data set, the research results are more effective for very young victims. The models and results provided in this study can provide a reference for public security departments to formulate policies in different regions, reduce the occurrence of illegal adoptions, or combat and intervene in the transfer of missing persons.

The following shortcomings emerged when designing and constructing the missing population prediction model. During data processing, we did not extract self-reported data on missing persons, which are written in natural language. This part of the analysis involved using descriptions of scenarios and transition processes recorded at the time of disappearance. When such content is extracted and analysed, more features can be obtained, and we were in turn able to establish a more complete prediction model.

Additionally, regarding data sources, missing persons' disappearance and arrival locations were derived from self-reported data. Mistakes made in the recall of missing persons led to the emergence of inaccuracies and uncertainty in the dataset and thus influenced the constructed model. Additionally, the missing person may move between multiple cities, but because there is no support of relevant data, the model cannot determine the trafficking chain. In the future, we hope to set up a model that can be used to track missing persons in real-time based on refined spatial data. With the expansion of data and the number of features examined, the prediction methods proposed in this study can be extended to real-time applications.

5. Conclusion

This study used the "Baby Coming Back Home" dataset to explore the China's missing population from a spatiotemporal perspective and predicted where missing populations may be located. Four independent models were obtained for location predictions. Moreover, an integrated prediction algorithm was constructed from model results using statistical methods, and city-level results with a recall rate of 82.99% were obtained.

In addition, using the completed RF models, we also examined what are the driving factors of destination. We find that the destinations of missing persons are determined by several factors, and locations of disappearance have the greatest impact on transfer processes.

This study is the first to predict the transfer locations and to explore factors that drive the whereabouts of missing persons based on an RF model. In the future, features of the original data will be more fully explored, and official data sources will be introduced to establish a complete model.

Data availability statement

The related data and codes that support the findings of this study are available with the identifiers at the private link: https://figshare.com/s/7a75a71dcb77ef8d20e1. All data were derived from the following resources available in the public domain: https://baobeihuijia.com/.

Author contributions

Qingfeng Guan, Yao Yao and Ye Hong designed the research; Yao Yao and Yifei Liu performed experiments and computational analysis, and drafted the manuscript; Ye Hong and Ruoyu Wang amended the manuscript; Xun Liang and Liangyang Dai contributed to the preparation of the experiments and computational analysis.

Funding

This work was supported by the National Key R&D Program of China (Grant No. 2019YFB2102903); National Natural Science Foundation of China (Grant No. 41801306, 41671408 and 41901332) and the Natural Science Foundation of Hubei Province (Grant No. 2017CFA041).

Declaration of Competing Interest

No conflict of interest exists in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part.

References

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249–2260.

Brewster, B., Ingle, T., & Rankin, G. (2014). Crawling open-source data for indicators of human trafficking. In *2014 IEEE/ACM 7th international conference on utility and cloud computing (UCC)* (pp. 714–719).

Chen, J., et al. (2013). Research on geographical environment unit division based on the method of natural breaks (Jenks). *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-4/W3, 4*, 47–50.

Chen, T., Zheng, J., & Jia, J. (2017). Shaping society: The changing logic of the relationship between state and society since the reform and open: An inspecting on guangdong's experiences. *Academic Research, 9*, 68–77.

Fajnzylber, P., Lederman, D., & Loayza, N. (2002). What causes violent crime? *European Economic Review, 46*(7), 1323–1357.

Fern, A., Ndez-Delgado, M., et al. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Kara, S. (2011). Designing more effective laws against human trafficking. *Journal of International Human Rights, 2*.

Laczko, F., & Ma, G. (2003). Developing better indicators of human trafficking. *Brown Journal of World Affairs, 10*(1).

Li, G., et al. (2017). Geographic characteristics of child trafficking crime in China. *Scientia Geographica Sinica, 37*(7), 1049–1058.

Liu, J. (2003). Consideration on the reformation of city -type units in China in new era. *Chinese Public Administration, 7*, 48–54.

Liu, J. (2014). Overeducation in China: Levels, trends and differentials. *Population Research, 38*(5), 8.

Liu, W., & Shi, E. (2016). Spatial pattern of population daily flow among cities based on ICT: A case study of "Baidu migration". *Acta Geographica Sinca, 71*(10), 1667–1679.

Liu, X., et al. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science, 31*(8), 1675–1696.
Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review, 94*(1), 155–189.
Ma, X., & Sun, C. (2011). The population fertility policy in 60 years of China. *Social Sciences of Beijing, 2011*(2), 46–52.
Ni, X., et al. (2008). The development process of comprehensive administration of social security since the reform and opening up. *Journal of Fujian Police Academy, 6*(106), 5–13.
Noort, S. P. V., et al. (2012). The role of weather on the relation between influenza and influenza-like illness. *Journal of Theoretical Biology, 298*(4), 131–137.
Obokata, T. (2005). Trafficking of human beings as a crime against humanity: Some implications for the international legal system. *International & Comparative Law Quarterly, 54*(2), 445–457.
Qiu, S., & Ma, Y. (2015). The essence of "missing" and the public security administration of missing persons. *Journal of Chinese People's Public Security University, 31*(5), 143–149.
Rudolph, A., & Schneider, F. (2014). International human trafficking: Measuring clandestinity by the structural equation approach. *Social Science Electronic Publishing, 51*(4), 374–378.
Schoon, I., & Polek, E. (2011). Teenage career aspirations and adult career attainment: The role of gender, social background and general cognitive ability. *International Journal of Behavioral Development, 35*(3), 210–217.
Shen, A., Antonopoulos, G. A., & Papanicolaou, G. (2013). China's stolen children: Internal child trafficking in the People's Republic of China. *Trends in Organized Crime, 16*(1), 31–48.
Stehman, S. (1996). Estimating the kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering and Remote Sensing, 62*(4), 401–407.
Sun, J., et al. (2012). The enlightenment of geographical theories construction from the first law of geography and its debates. *Geographical Research, 119*(22), 1749–1763.
Svetnik, V., et al. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947.
Thapa, R. B., & Murayama, Y. (2009). Urban mapping, accuracy, \& image classification: A comparison of multiple approaches in Tsukuba City, Japan. *Applied Geography, 29*(1), 135–144.
Tian, X. (2012). Baby abandoning and adoption: The collision and coupling between family planning and fertility culture in the village - a microscopic explanation to the practical logic of rural family planning policy in Gannan in 1990s. *Youth Studies, 1*, 38–49.
Wang, T. (2016). Immigrants, Hukou System and criminal offense. *Population Research, 40*(2), 63–74.
Wang, Y. (2015). *Study of th imapct of China's openness on its crime rate*. Nanjing University.
Wang, Z., et al. (2018). Child-trafficking networks of illegal adoption in China. *Nature Sustainability, 1*(5), 254–260.
Xu, J., et al. (2005). Spatial and temporal scale analysis on the spatial and temporal scale analysis on the regional economic disparities in China. *Geographical Research, 24*(1), 57–68.
Yao, Y., et al. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS, 22*(2), 561–581.
Yao, Y., et al. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science*, 1–22.
Zhang, A., Ballas, N., & Pineau, J. (2018). *A dissection of overfitting and generalization in continuous reinforcement learning*. arXiv:1806.07937.
Zhao, Z., & Wang, S. (2017). A spatial-temporal study of inter-provincial migration pattern during Chinese spring festival travel rush. *Population Research, 41*(3), 101–112.
Zheng, J., Wu, H., & Wang, B. (2014). Summary of the climate in Guangdong Province in 2013 and analysis of the causation of climate abnormality. *Guangdong Meteorology, 36*(1), 26–29.