Quantifying Human Mobility Using

The Longest Distance Traveled

by
Yanjie Huang

A thesis submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science

Baltimore, Maryland

April 2014

Abstract

Background: Population movement has a dramatic impact on infectious disease epidemiology. Human mobility data is increasingly being used to model pathogen dispersion, but it is difficult to study long-distance movement of humans.

Objectives: Develop models of long-distance human travel based on questionnaire data describing the longest distance traveled by someone in a household over one week, one month, 6 months, and 1 year.

Methods and Analysis: Mathematical models were generated based on the daily travel and N-day routine hypotheses. The parameters were estimated through maximum likelihood method, and Bayesian information criteria (BIC) was used to assess the goodness of fit.

Results and Conclusions: Household location was an important factor shaping human mobility. A routine-travel cycle consisting of 11-13 days provided the best fit (BIC = 29802). Daily human movement and human movement overall may be best explained by routines that are repeated biweekly or perhaps longer.

Advisors and Readers:

Dr. Derek A. Cummings

Dr. David Smith

Acknowledgements

This thesis represents a milestone in my life. My experience at Johns Hopkins University School of Public Health has been nothing short of amazing. Since the first day, I have been given unique opportunities to work with remarkable persons who I wish to acknowledge.

First and foremost I would like to acknowledge my two advisors, **Dr. Derek Cummings** and **Dr. David Smith**. They are amazing persons in too many ways. Throughout the past two years, under their guidance, I have experienced a fantastic journey in the field of infectious disease dynamic. It has been an honor for me to be their student.

Dr. Derek Cummings has been my advisor and guiding me since February 2013. I am grateful for my embarkation on his project FluScape after our first meeting. His steadfast integrity and immense knowledge motivate me to complete this thesis. He has the great instinct for understanding and guiding my research. From him, I have learnt the vital skill to understand the real world from what we see in the data. Dr. David Smith, my original academic advisor, he has guided me since my first day here. He gave me the freedom to shape my study, and encouraged me to challenge myself with tough goals and provided guidance on how to reach those goals. Without him, I would not be able to understand the art of describing and explaining real-time phenomenon through math. I truly appreciate his insightful support during his two-year honest mentorship. Dr. Derek Cummings and Dr. David Smith have been such important persons for both my personal and academic development. I owe a great debt of gratitude to them.

Furthermore, I would like say thank you to **Dr. Justin Lessler** for guiding me through the entire process. Without him, the thesis would not have been possible. He generously helped me to come up with fantastic ideas and guided me over the past year. He is such a nice, wise and patient person to work with that I remain greatly appreciative. And thank you to **Dr. Ciprian Crainiceanu**, with whom I was able to think deeply about the research question. Also thank you to **Dr. Henrik Salje** and **Shaojie Chen**, who have helped with constructive suggestions.

Lastly, to my family and friends I am fully grateful for your support. Specially **Long Zhang**, who has been part of my life for five years. During my most struggling days, he gave me the moral support and unlimited love. I owe so much thanks to him. Also I wish to acknowledge my friends **Alpana Kaushiva** and **Yingjie Wen** for their support and understanding during the completion period of my thesis.

Contents

List of Tables

List of Figures

Introduction

Population movement has a dramatic impact on the spread of infectious diseases. Since the diffusion of plagues in 1970s[1], the past decades have witnessed a remarkable development of understanding of human mobility and its contribution to the spatial spread of various infectious diseases, including plagues[1], malaria[2-5], influenza[6-8], SARS[9,10], cholera[11], HIV and other sexually transmitted diseases[12-14]. In particular, this impact has been enhanced by the increasing globalization and transportation infrastructure. Globalization enables the connection between people located at any points on earth, promoting the dispersal of infectious disease on a large scale. Additionally, significant reductions in transportation costs have altered human travel behaviors. Frequent and longer trips increase the odds of pathogens exploiting larger pools of humans and/or animals, supporting higher levels of endemic transmission. Understanding patterns and the magnitude of human mobility will provide important tools to characterize the potential spread of infectious diseases and help determine optimal control measures.

Human mobility is an extremely complex process. One of the challenges in this field is how to collect mobility data effectively to reveal key human travel patterns. Among all attempts to explore the human mobility pattern, direct observation of movement is the most powerful to capture the overall spatial diffusion. However, tracing actual movement paths consumes many resources, and there are limits to what can be learned about the drivers of human travel from de-identified aggregate data. People have used bank notes[15,16] and smart cell phones[17-21] to estimate the human spatial dispersal. One of the alternative approaches to recapture the redistribution is to record the long distance movements. Long-distance dispersal, of long interest in ecology, plays a critical role in seed dispersal and species invasion[22-25]. For humans, long-range air flights have been studied to characterize spatial movements of humans at large scales[15,26-29]. In this study, we want to look into the properties of human travel behavior at smaller scales by characterizing survey data that asks respondents in southern China about travel in the past week, month, half year and year.

Mathematical models have been key tools in understanding human movement. Though increasing efforts have been directed towards building models of human movement, multiple models exist, and it is not clear which models perform best in fitting empirical data on movement of humans. The random and unforeseeable appearance of human travel leads numerous studies to describe the human travel trajectory as a stochastic process like a random walk or Levy flight[16,30,31], which assumes that there is no difference between the movements of humans and bacterium. However, other models begin with the premise that economic or social forces like work opportunities drive human travel to respond to the spatial distribution of humans. These models include the gravity model[32-36], logit model[37,38] and radiation model[39]. The gravity model relies on the hypothesis that the traffic volume between two locations is related to their population size and their distance. Analogous to the gravity model, the logit model introduces the utility function into the trip decision-making. Additionally, the radiation model is based on the assumption that the distance individuals' travel responds to the number of opportunities they pass when traveling a given distance.

Here, we take an empirical approach to describing human movement and its association with individual, household and community scale factors. We propose a mathematical framework that is based on a hypothesis that human travel is routine with certain temporal cycles and is associated with individual, household and community level characteristics. We fit multiple probability distributions to observed travel patterns among individuals in the Fluscape study, a longitudinal study of human movement and influenza in southern China[40,41]. These results provide an empirical basis upon which mechanistic models like those described above can be built.

As far as we know, our study is the first attempt to explore the modalities and properties of human travel through the furthest travel distance within various lengths of time periods. In order to assess the regularity of human travel and its association with multiple covariates, we built multiple models of varying structure and included multiple candidate covariates and then tested the consistency of these models with data on human travel collected in southern China.

Materials and Methods

Study Area and Study Design

All the data come from the Fluscape project conducted in Guangdong Province, China. The Fluscape project is a longitudinal study using validated questionnaires on household structure, travel and social contacts to gather data every year. We used data collected from December 4, 2009 to November 22, 2011. In total, there were 1096 households selected from 40 randomly selected study locations in a fan-shaped area (centered at Guangzhou and extending to the northeast, Figure 1). All the selected households were administered and completed questionnaires that asked participants for information about their household (household's location, household size, monthly household income and the ownership of motor vehicle) and the household-level longest travels undertaken within the past 7 days, 30 days, 6 months and 1 year (including identification of the name of the destination). The specific question that heads of households were asked was "Name the furthest location that anyone living in this household has traveled from home in the past X days". Latitude and longitudes of each of the named locations was determined using Google Earth [42].

Analysis

The great circle distance was used to calculate the distance between points over the earth's surface. We fit multiple models to the reported furthest travel distance traveled by any household member. Parameters were estimated through maximum likelihood. To avoid over-fitting, we evaluated the goodness of fit among models through Bayesian information criteria (BIC).

We tested two hypotheses:

Hypothesis 1: Human travel follows a fixed daily pattern, and the pattern varies with the type of the day (e.g. national holidays, school holidays and ordinary days), as well as by the spatial location and characteristics of households.

Hypothesis 2: Human travel follows a fixed daily pattern as well as a regular routine pattern with a cycle consisting of some number of days longer than one (to be estimated). We assessed these hypotheses by fitting multiple probability distributions describing travel distance to the observed data where separate models were fit to different stratifications of the data by covariates of interest. A description of each of the models appears below.

Hypothesis 1. Fixed Daily Travel Pattern

For each household, the daily travel-distance within each day is assumed to be independent and identically distributed (i.i.d). We assume the probability distribution of daily travel-distance follows a lognormal distribution.

$d_i \sim Lognormal(\mu, \sigma)$:

$$F(d) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log(d) - \mu}{\sigma\sqrt{2}}\right)\right] = \phi\left(\frac{\log(d) - \mu}{\sigma}\right)$$

The cumulative probability distribution:

$Pr(Mn < d) = Pr(d_1 < d, d_2 < d, \cdots, d_n < d)$

$$= Pr(d_1 < d) \times \cdots \times Pr(d_n < d)$$

$$= F(d) \times \cdots \times F(d)$$

$$= (F(d))^n$$

The basic form (Model 1.1):

$$Pr(M_7 < d) = \left(F(d)\right)^7$$

$$Pr(M_{30} < d) = \left(F(d)\right)^{30}$$

$$Pr(M_{183} < d) = \left(F(d)\right)^{183}$$

$$\Pr(M_{365} < d) = \left(F(d)\right)^{365}$$

ϕ: *the integral of the standard normal distribution*

d_i: *the household daily travelled distance for day i*

Mn: *the household maximum travelled distance for n days*

$Mn = max\{d_1, \dots, d_n\}$

$F(d)$: *the cumulative probability distribution for* d_i

Considering the impact of temporal heterogeneity, spatial heterogeneity, and household heterogeneity, we built Model 1.2 – 1.5 by stratifying the population by the covariates as indicated in table X and fitting separate models to each strata.

Hypothesis 2. Fixed Daily Travel + Regular Routine Travel Pattern

For each household, the daily travel-distribution within each day and the routine travel distance within each cycle are assumed to be independent and identically distributed (i.i.d). We also assume that both of the probability distributions for daily travel-distance and routine travel pattern follow a lognormal distribution.

$d_i \sim Lognormal(\mu_1, \sigma_1)$:

$$F(d) = \frac{1}{2}\left[1 + erf\left(\frac{log(d) - \mu_1}{\sigma_1\sqrt{2}}\right)\right] = \phi\left(\frac{log(d) - \mu_1}{\sigma_1}\right)$$

$D_i \sim Lognormal(\mu_2, \sigma_2)$:

$$F(D) = \frac{1}{2}\left[1 + erf\left(\frac{log(D) - \mu_2}{\sigma_2\sqrt{2}}\right)\right] = \phi\left(\frac{log(d) - \mu_2}{\sigma_2}\right)$$

The cumulative probability distribution:

$$Pr(Mn < d) = Pr\left(d_1 < d, d_2 < d, \cdots, d_n < d, D_1 < d, D_2 < d, \cdots, D_{[\frac{n}{cycle}]} < d\right)$$

$$= Pr(d_1 < d) \times \cdots \times Pr(d_n < d) \times Pr(D_1 < d) \times \cdots \times Pr\left(D_{[\frac{n}{cycle}]} < d\right)$$

$$= F(d) \times \cdots \times F(d) \times F(D) \times \cdots \times F(D)$$

$$= (F(d))^n \times (F(D))^{[\frac{n}{cycle}]}$$

The basic form (Model 2.1):

$$Pr(M_7 < d) = \left(F(d)\right)^7 \times \left(F(D)\right)^{[\frac{7}{cycle}]}$$

$$Pr(M_{30} < d) = \left(F(d)\right)^{30} \times \left(F(D)\right)^{[\frac{30}{cycle}]}$$

$$Pr(M_{183} < d) = \left(F(d)\right)^{183} \times \left(F(D)\right)^{[\frac{183}{cycle}]}$$

$$Pr(M_{365} < d) = \left(F(d)\right)^{365} \times \left(F(D)\right)^{[\frac{365}{cycle}]}$$

Mn: *the household maximum travelled distance for n days*

$Mn = max\{d_1, \dots, d_n, D_1, \dots, D_m\}$

$F(d)$: $the\ cumulative\ probability\ distribution\ for\ d_i$

$F(D)$: $the\ cumulative\ probability\ distribution\ for\ D_i$

$cycle$: $the\ integer\ length\ of\ the\ regular\ routine\ travel\ cycle$, $cycle =$

$1, 2, 3, \dots, 365$

We used the following algorithm to estimate the cycle with the smallest minimum log likelihood ($-ln\mathcal{L}$)

1. Start at $cycle = 1$

2. Estimate the minimum $-ln\mathcal{L}$ for $cycle$

3. Repeat step 2 with updated value of $cycle$ from 2 to 365

4. Find the value of $cycle$ with the smallest minimum $-ln\mathcal{L}$

5. Estimate other parameters based on $\widehat{cycle}$

Considering the impact of temporal heterogeneity, spatial heterogeneity, and household heterogeneity, we built Model 2.2 – 2.5 with stratification.

Data with missing values in any of the covariates required for any of the models lead to exclusion from the analysis. After data cleaning, there were 715 (65.24%) records left. All the analyses were performed using the statistical R package (3.0.2) (R Core Team, Vienna, Austria; http://www.R-project.org).

Results and Conclusion

Characteristics of the furthest travel distance

Before the model construction, we explored the characteristics of the human furthest travel. The median of the furthest travel distance within 7 days, 30 days, 6 months and 1 year is 6.64 km, 14.64 km, 32.32 km, and 45.01 km respectively. A considerable fraction of the longest travel is reported to occur within the neighborhood of the household location (the proportion of the longest travel distances ≤ 10 km: within 7 days, 63.4%; within 30 days, 41.7%; within 6 months, 22.8%; within 1 year, 16.6%). Meanwhile, the proportion of the furthest travel distance that is above 100 km is also notable (the furthest travel distance >100 km: within 7 days, 4.3%; within 30 days, 10.6%; within 6 months, 24.8%; within 1 year, 31.5%).

Figure 2 shows the spatial distribution of the furthest-travel destination. As the temporal observation window extends, the geographical boundary of the furthest travel stretches far away from Guangzhou City, with furthest-trips reaching places with distance > 10000 km within 30 days. Given longer time-window, people tend to take the furthest-trips to other cities (out of Guangzhou City: within 7 days, 10.5%; within 30 days, 23.2%; within 6 months, 43.8%; within 1 year, 50.3%), other provinces (out of Guangdong City: within 7 days, 2.8%; within 30 days, 6.3%; within 6 months, 14.7%; within 1 year, 22.2%), and even other countries (out of China: within 7 days, 0.0%; within 30 days, 0.1%; within 6 months, 0.4%; within 1 year, 1.1%).

As expected, the distribution of the furthest travel distance shifts to right as the time interval increases (Figure 3-A). Beside the main peak for each distribution, there are two additional peaks located at around 1 km and 310 km ($10^{2.5}$ km). In order to further understand the properties of the furthest travel distance, we generated the graph for complementary cumulative probability (1 – the cumulative probability with distance) (Figure 3-B). The decay rate before 100 km within 7 days is the greatest one among the four time windows. We find that the complementary cumulative probability distribution for both 6 months and 1 year are similar, indicating that majority of the furthest travels within 1 year are possibly achieved within only 6 months.

There is heterogeneity in travel distance among different time, spatial areas and households. The shapes of furthest travel distance within 4 time windows vary over different months (Figure 4-A). Considering the fact that people were asked about travel at different times in different sampled locations, the observed differences by location could be attributable to temporal differences in travel patterns. Figure 4-B shows patterns for 6 districts offered further evidence of the spatial heterogeneity. The households in rural areas (Luogang, Conghua, and Zengcheng) tend to achieve longer furthest-distance trips within short terms (7 days and 1 month), and those in urban area (Yuexiu and Tianhe) are more likely to travel long furthest-trips within long terms (6 months and 1 year). Different from them, the travel behavior for households in suburban area (Baiyun) shows a unique pattern. For households located in this area, the furthest trips within 7 days are longer than the trips taken by households in other areas. However, for the furthest trips within longer time intervals (6 months and 1 year), this feature no longer exists. The heterogeneity in the impacts of different income levels on furthest trips is obvious (Figure 4-C). Across all time windows, the richer households are more likely to take longer furthest trips.

Based on the two different hypotheses, there are two main series of models in our analysis: the daily travel model, and the daily travel and regular routine model.

Performance of daily travel model

We have fitted various distributions via Maximum Likelihood for the null model. In terms of characterizing the entire furthest travelled distance within four time windows, the lognormal distribution does a better job offering the best fit with BIC = 31096.61 than the Weibull distribution and Gamma distribution. The estimated median distance for daily travel in Model 1.1 is 1.4 meters (95% CI: 1.1 ~ 1.7 meters). After accounting for temporal, spatial and household heterogeneity, the predictive performance of daily travel distance distribution is improved (Table 1). We note that the daily travel model with stratification by household income and districts (Model 1.5.3) is the best one to capture the overall distribution of furthest travel distance within 7 days, 30 days, 6 months and 1 year (BIC = 30566.62). The estimates of parameters for Model 1.5.3 are presented in Table 3.

These estimation results are problematic. The estimates of median distance for daily travel have a range from 0.1 m to 1161.8 m. And the standard deviation for daily-travel distance on the log scale ranges from 1.29 to 4.84. They are much lower than what we observe in our daily life. However, besides those numbers, they show two major features. One of them is that households in the rural and suburban areas take longer daily trips. The other one is that, besides Baiyun District, households with a monthly household income higher than 5000 RMB tend to have the longest daily trips. Overall, those models are not able to facilitate good fittings. Based on the poor performance of this model series, the first hypothesis is proved to be not close enough to explain our data.

Performance of daily travel and regular routine model

In order to verify the second hypothesis that there is a regular routine travel pattern with certain cycle, we included a term representing the regular routine into the model.

Before the model comparison, the first step is to estimate the cycle length for the regular routine. With restricting the cycle as an integer ranging from 1 to 365, we are able to assess the maximum value of likelihood for each potential cycle length within the defined range. Then, we are able to find out the cycle length with the largest maximum likelihood. After repeating this procedure in different sub-datasets, surprisingly, we note a certain pattern indicating there is a routine travel cycle consisting of 11 to 13 days. By including a routine travel with such a cycle, the impact on the reduction of the negative log-likelihood value is apparent. Additionally, this pattern is stably identical through all stratified subgroups (Figure 5). This phenomenon supports our Hypothesis 2 that for human travel behavior, beside the daily travel pattern, there is a regular routine travel with a cycle consisting of several certain days. And the promising length of the regular routine trips is around 12 days. We believe that there are two distinct mobility behaviors for human travel activity, and this regularity pattern could be quite stable and universal.

Compared with the fixed daily travel model, the model with regular routine provides much better fitting (Table 2). Even for the null model, the regular routine significantly improves the model performance by reducing BIC from 31096.61 (Model 1.1) to 30515.66 (Model 2.1). Additionally, the null model (Model 2.1) in this series beats the best model (Model 1.5.3) in the previous series for the fixed daily travel model. The improvement was observed across all the sub-models with the regular routine. Again, it provides convincing evidence that besides daily travel behavior, there is a regular routine travel pattern.

Among all the models, the one with stratification by household locations (6 districts: Yuexiu, Tinahe, Baiyun, Luogang, Conghua, and Zengcheng) provides the best fitting (BIC = 29802.18, Figure 6). The results suggest that, both for the daily travel and the regular routine travel, the spatial heterogeneity is the most important factor in explaining the variation in travel behavior. The estimates of median distance for daily travel for households in 6 different districts are 386.7 meters (Yuexiu), 173.8 meters (Tianhe), 726.1 meters (Baiyun), 142.3 meters (Luogang), 140.9 meters (Conghua), and 143.7 meters (Zengcheng). And the estimates of median distance for each regular routine travel are 45 meters (Yuexiu), 316.6 meters (Tianhe), 537.9 meters (Baiyun), 663.7 meters (Luogang), 860.7 meters (Conghua), and 657 meters (Zengcheng). Except for Yuexiu District, the median daily travel distance is overall shorter than the regular routine travel.

And in general, the median daily trips for households in urban and suburban areas are longer than the median daily trips in rural area. However, this pattern of regular routine travel shows an entirely distinct pattern. Households in rural areas tend to take longer regular routine travel, and the length of routine cycle is slightly longer. In other words, compared with households located in urban areas, those households in rural areas are more likely to take short daily trips more frequently, and take longer regular routine travel less frequently. The estimates of parameters for Model 2.3.2 are show in Table 4.

Discussion

Most of the existing research on human mobility tests complex theory or relies on local socioeconomic information. The cost of taking such an approach may be a loss of ability to reveal the fundamental characteristics of human travel. Here, we have analyzed questionnaire data describing the longest distance moved by families over various periods of time. Based on analysis of that data, the properties of extreme value statistics, and a novel mathematical framework allowing us to test plausible hypotheses, at least some new aspects of human mobility.

Our study exhibits two vital properties of the human travel. First of all, besides daily travel, our analysis suggests there is a clearly regular routine travel with a cycle consisting of 11 to 13 days. This feature is fairly stable, and it could reveal a universal aspect of human movement habits and behavior. In the real world, human daily trips are shaped because of various purposes like jobs, schools, and business. Over slightly longer periods, there may be predictable breaks from a regular routine, such as visiting relatives, friends, and an escape from daily life. These "occasional" events, which may dominate the reasons for long-distance travel, may be of great importance for the spread of infectious diseases.

We also learn that the spatial heterogeneity is the factor that best explains variation in the travel distances. In our study, the daily travel and regular routine travel show different shapes for different areas. Households in urban and suburban areas tend to take longer daily trips. For households in rural areas, longer routine trips appear to occur on this longer cycle.

It seems reasonable that human mobility behaviors are hugely shaped and regulated by where we live. From a large scale, the transportation system and the unique resources found in a city largely determine where we are going and how to go there. From the individual's perspective, social economical status and health status are closely related with where we live and mobility patterns.

However, our results also indicate that our models have limited ability to describe short-term travel. Based on Figure 6-A and 6-B, our model clearly overestimates the probability of furthest travel with distance from 10 meters to 1000 meters, but underestimate the probability of the furthest travel distance with 10000 to 100000 meters. It fits better for long-term furthest travel, especially for 6-month and 1-year time period. The result shows the potential in-homogeneity of travel distance within short term and long term.

One of the potential explanations is that the i.i.d assumptions we made may be unrealistic for the human travel. There could be an auto-regression relationship that the travel distance we made on a certain day is dependent upon the trips we took in the previous days. Another possible reason is that the distribution describing regular travel routines could be different from lognormal distribution. It may also be true that travel cycles at monthly or annual time scales.

Additionally, the parameter estimates based on maximum likelihood that results in assigning heavier weight for long-term travel is another reason that the estimate is more accurate for long term. This is one of the limitations of this study. During the procedure of maximum likelihood estimate, it autumnally assigns heavier weight by giving the term days as 183 (6 months) or 365 (1 year).

Another limitation of our study is the lacking of prediction of travel direction. The shapes of the density are not ideally smoothing and standard distributed (Figure 3-A). Our model has limited ability to capture the other peaks located at around 1 km and 310 km. Causing by higher probability of travelling to locations with the certain length, it reveals the disparate spatial distribution of travelled destinations. Our studies in the future will try to combine with spatial social network into our framework.

Despite the limitations, our study is useful for the field of epidemiology. When modeling geographic spread of infectious disease and constructing the early warming system, it is important to take the spatial heterogeneity into consideration. Some places are more mobile than others. Furthermore, travel patterns such as daily trips and regular routine travel could have different roles in the infectious disease spread. The study of longest distance traveled could be a promising starting point for understanding general patterns and causes of individual travel patterns and exploring their effects on disease transmission.

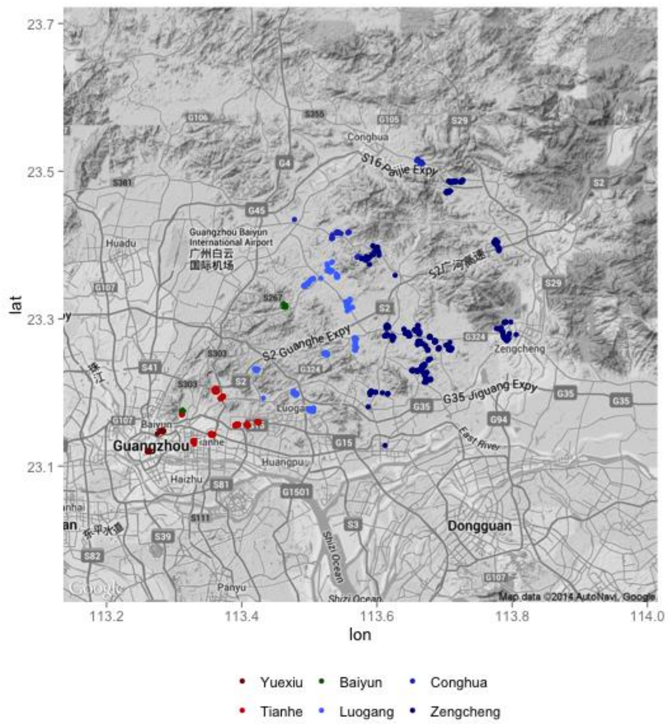

Figure 1. Spatial Distribution of Sampled Households in Guangzhou, China

Our samples were located in 6 different districts in Guangzhou. Yuexiu District (dark red) and Tianhe District (red) belong to urban area. Baiyun District (green) is located in the suburban area. Luogang District (light blue), Conghua District (blue) and Zengcheng (dark blue) are in the rural area. The number of sampled households for Yuexiu, Tianhe, Baiyun, Luogang, Conghua, and Zengcheng are 34 (4.78%), 133 (18.68%), 33 (4.63%), 154 (21.63%), 41 (5.76%) and 317 (44.52%) respectively.

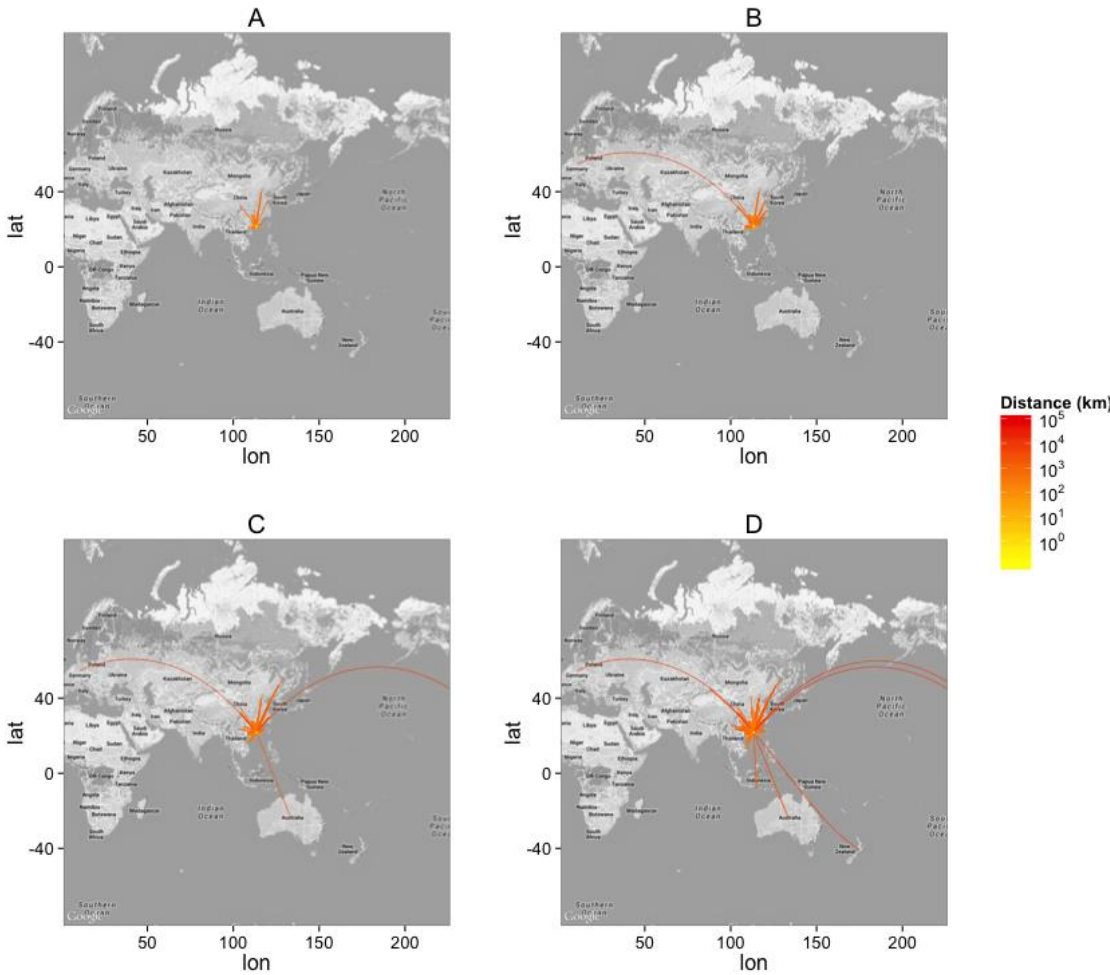

Figure 2. Spatial Distribution of the Destination for the Furthest-Travel within 7 Days, 30 Days, 6 Months and 1 Year

(A) Trips within 7 days; (B) Trips within 30 days; (C) Trips within 6 months; (D) Trips within 1 year.

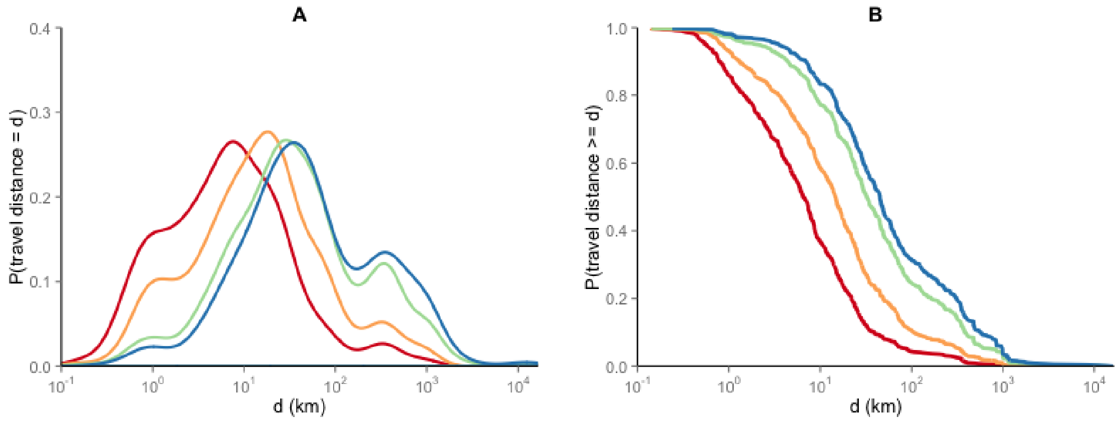

Figure 3. Probability Distribution for the Furthest Travel Distance

(A) Probability distribution for furthest travel distance by 7 days (red), 30 days (orange), 6 months (green) and 1 year (blue); (B) Complementary cumulative probability distribution for furthest travel distance by 7 days, 30 days, 6 months and 1 year.

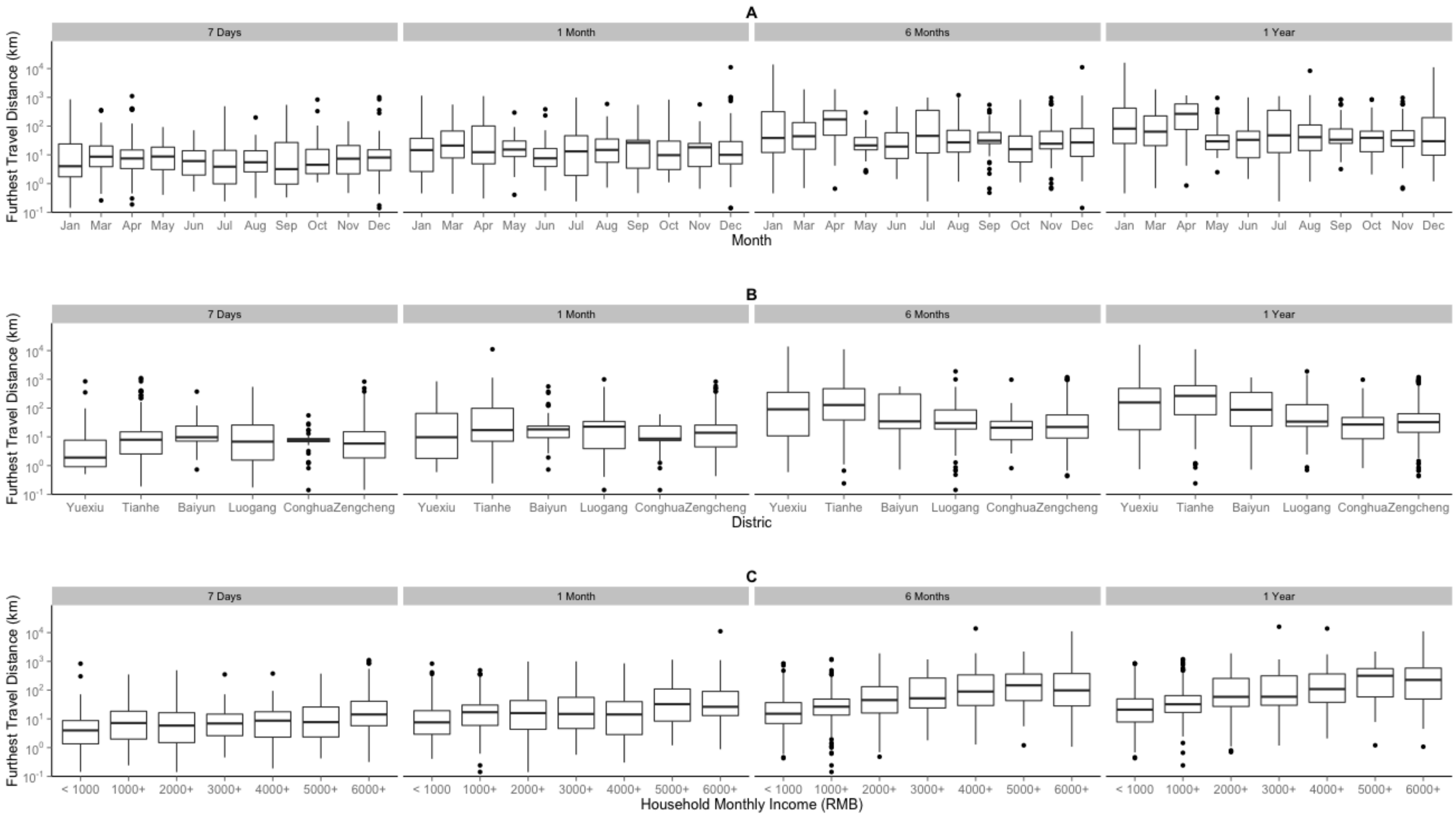

Figure 4. Distribution of Furthest Travel Distance with Stratification by Investigated Date, Household Location and Household Income

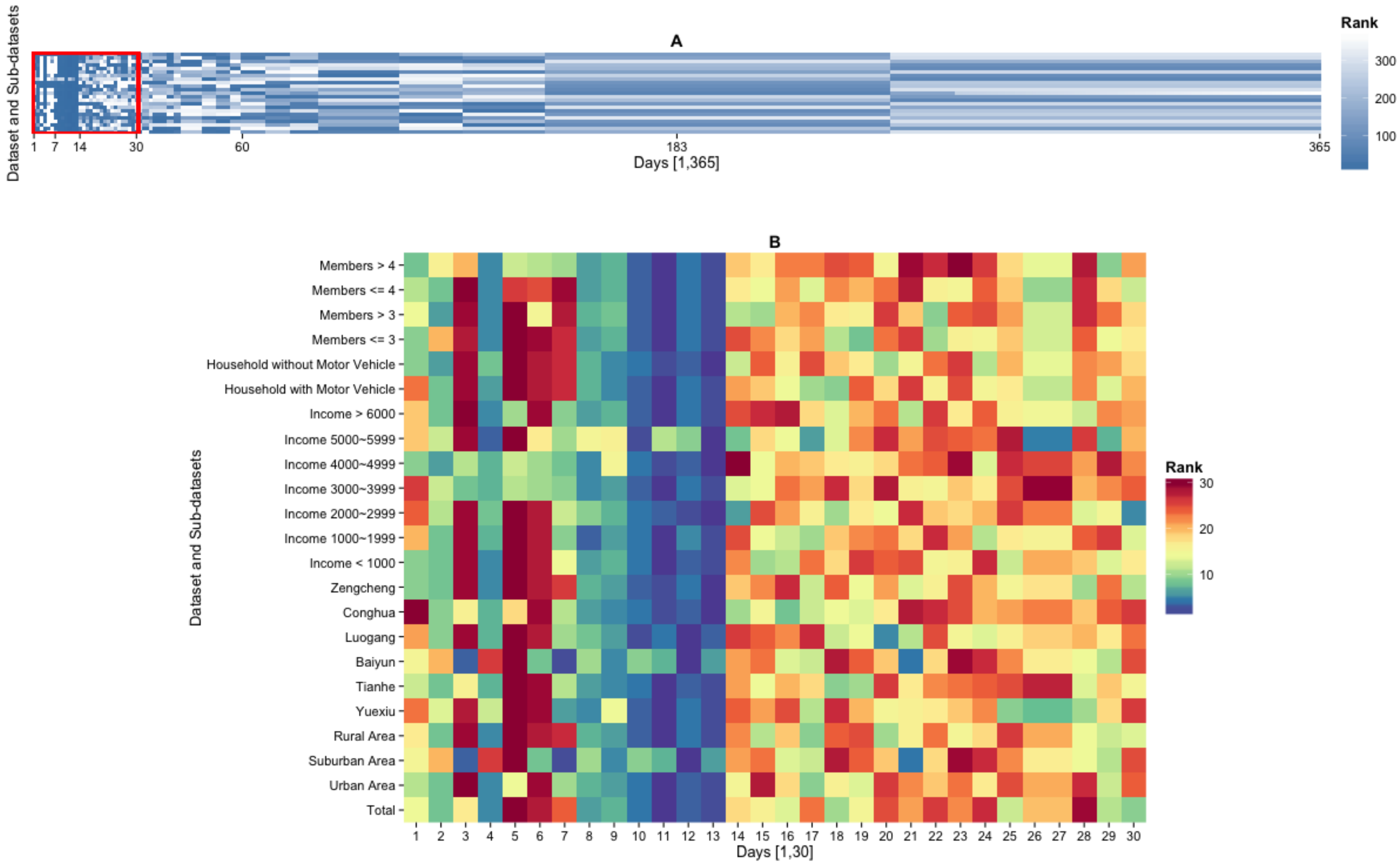

Figure 5. The Rank of Negative Log-Likelihood Value by Different Cycles Settings

With restricted the cycle as integer ranging from 1 to 365, the cycle of the regular routine was estimated through Maximum Likelihood. We ranked the negative likelihood values. (A) The smallest ranks exist between 7 days to 14 days. (B) The pattern becomes obvious by zooming in the area in red box in (A: 1~30 days). This evidence supports Hypothesis 2.

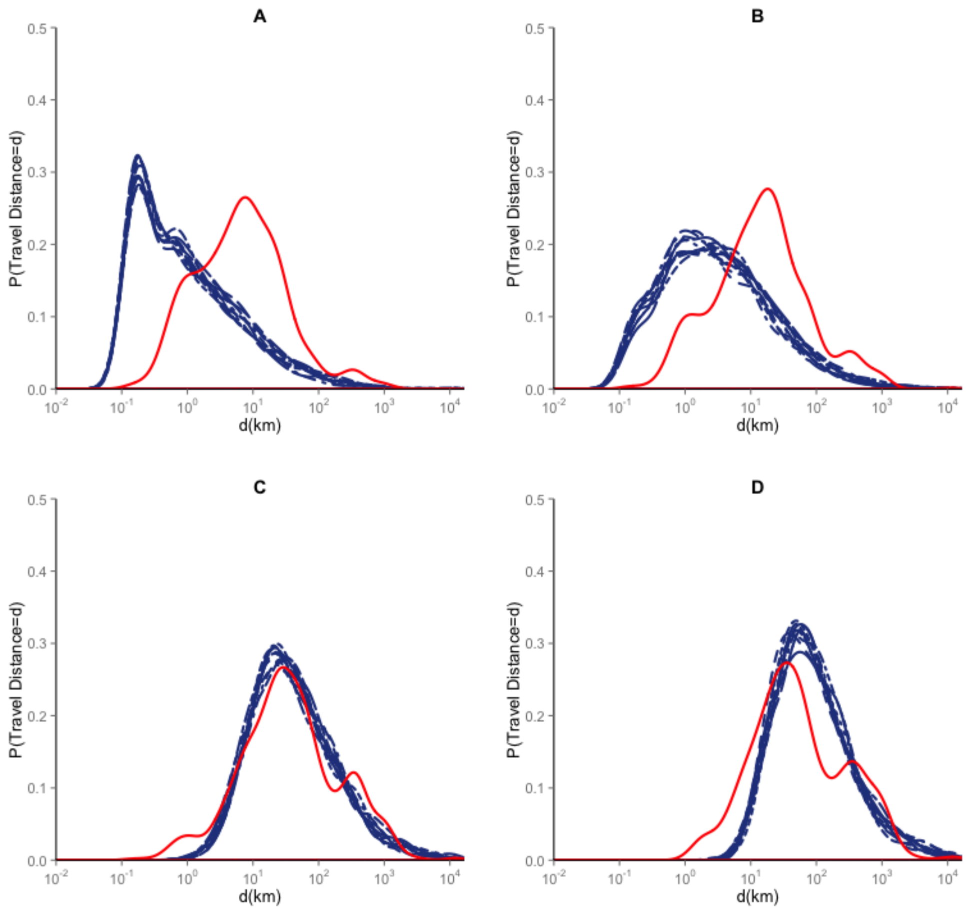

Figure 6. The Density of Observed Furthest Trips and Predicted Values based on Model 2.3.2

(A) Trips within 7 days; (B) Trips within 30 days; (C) Trips within 6 months; (D) Trips within 1 year. Solid lines in red are the observed data. Dashed lines in blue are the density distribution for simulated data based on Model 2.3.2, which includes the terms regarding to 6 districts.

Table 1. Summary of the Fixed Daily Travel Models

Models		Number of Parameters	$-ln\mathcal{L}$	AIC	BIC
Model 1.1 Null model		2	15540.35	31084.70	31096.61
Model 1.2 Temporal Heterogeneity					
Model 1.2.1	National Holiday, Non-Holiday	4	15527.25	31062.50	31086.32
Model 1.2.2	National Holiday + School Holiday, Non-Holiday	4	15529.94	31067.88	31091.70
Model 1.2.3	National Holiday, School Holiday, Non-Holiday	6	15517.35	31046.70	**31082.43**
Model 1.2.4	National Holiday, School Holiday (Winter), School Holiday (Summer), Non-Holiday	8	15511.07	**31038.14**	31085.77
Model 1.3 Spatial Heterogeneity					
Model 1.3.1	Urban, Suburban, Rural	6	15406.61	30825.22	30860.95
Model 1.3.2	6 Districts	12	15373.12	**30770.24**	**30841.69**
Model 1.4 Household Heterogeneity					
Model 1.4.1	Household Income (7 levels)	14	15277.43	**30582.86**	30666.22
Model 1.4.2	Household Income (3 levels)	6	15306.93	30625.86	**30661.59**
Model 1.4.3	Motor vehicle	4	15459.96	30927.92	30951.74
Model 1.4.4	Household size 1 (Members > 3, Members ≤ 3)	4	15462.51	30933.02	30956.84
Model 1.4.5	Household size 2 (Members > 4, Members ≤ 4)	4	15478.41	30964.82	30988.64
Model 1.5 Combinations					
Model 1.5.1	Temporal Heterogeneity (1.2.3), Household heterogeneity (1.4.2)	18	15285.03	30606.06	30713.24
Model 1.5.2	Temporal Heterogeneity (1.2.3), Spatial Heterogeneity (1.3.2)	36	15303.34	30678.68	30893.04
Model 1.5.3	Spatial Heterogeneity (1.3.2), Household Heterogeneity (1.4.2)	36	15140.13	30352.26	**30566.62**
Model 1.5.4	Temporal Heterogeneity, Spatial Heterogeneity, Household Heterogeneity	108	15035.27	30286.54	30929.61

National public holidays in China include New Year's Day, Chinese New Year, QingMing Festival, Labor Day, Dragon Boat Festival, Mid-Autumn Days, and National Days. School holidays consist of winter and summer vacations.

Table 2. Summary of the Daily Travel + Regular Routine Models

Models		Number of Parameters	$-ln\mathcal{L}$	AIC	BIC
Model 2.1 Null model		4	15241.92	30491.84	30515.66
Model 2.2 Temporal Heterogeneity					
Model 2.2.1	National Holiday, Non-Holiday	6	15241.43	30494.86	30530.59
Model 2.2.2	National Holiday + School Holiday, Non-Holiday	6	15241.26	30494.52	**30530.25**
Model 2.2.3	National Holiday, School Holiday, Non-Holiday	8	15240.05	30496.10	30543.73
Model 2.2.4	National Holiday, School Holiday (Winter), School Holiday (Summer), Non-Holiday	10	15229.85	30479.70	30539.24
Model 2.3 Spatial Heterogeneity					
Model 2.3.1	Urban, Suburban, Rural	12	14944.07	29912.14	29983.59
Model 2.3.2	6 Districts	24	14805.64	**29659.28**	**29802.18**
Model 2.4 Household Heterogeneity					
Model 2.4.1	Household Income (7 levels)	28	14871.39	**29798.78**	**29965.50**
Model 2.4.2	Household Income (3 levels)	12	14980.70	29985.40	30056.85
Model 2.4.3	Motor vehicle	8	15169.62	30355.24	30402.87
Model 2.4.4	Household size 1 (Members > 3, Members ≤ 3)	8	15181.68	30379.36	30426.99
Model 2.4.5	Household size 2 (Members > 4, Members ≤ 4)	8	15170.21	30356.42	30404.05
Model 2.5 Combinations					
Model 2.5.1	Temporal Heterogeneity (2.2.2), Household heterogeneity (2.4.1)	35	14864.11	29798.22	30006.62
Model 2.5.2	Temporal Heterogeneity (2.2.2), Spatial Heterogeneity (2.3.2)	30	14797.61	29655.22	29833.85
Model 2.5.3	Spatial Heterogeneity (2.3.2), Household heterogeneity (2.4.1)	----	----	----	----
Model 2.5.4	Temporal Heterogeneity, Spatial Heterogeneity, Household Heterogeneity	----	----	----	----

National public holidays in China include New Year's Day, Chinese New Year, QingMing Festival, Labor Day, Dragon Boat Festival, Mid-Autumn Days, and National Days. School holidays consist of winter and summer vacations.
For Model 2.5.3 and 2.5.4, there are 42 sub-datasets. 12 of them include only around 10 observations. The small sample data sets in general provide less reliable bases for estimating models, so we did not present results here.

Table 3. Summary Table for Model 1.5.3

District	Monthly Household Income	μ (sd)
Yuexiu	< 1000 RMB	-9.10 (4.05)
	1000~4999 RMB	-9.78 (4.87)
	>5000 RMB	-8.38 (4.78)
Tianhe	< 1000 RMB	-10.10 (4.84)
	1000~4999 RMB	-8.47 (4.76)
	>5000 RMB	-5.84 (4.03)
Baiyun	< 1000 RMB	-10.29 (4.80)
	1000~4999 RMB	-1.75 (1.97)
	>5000 RMB	-4.31 (3.35)
Luogang	< 1000 RMB	-5.54 (3.24)
	1000~4999 RMB	-5.85 (3.52)
	>5000 RMB	-4.91 (3.48)
Conghua	< 1000 RMB	-4.30 (2.41)
	1000~4999 RMB	-3.00 (2.39)
	>5000 RMB	0.15 (1.29)
Zengcheng	< 1000 RMB	-5.84 (3.18)
	1000~4999 RMB	-5.12 (3.17)
	>5000 RMB	-4.69 (3.22)

Table 4. Summary Table for Model 2.3.2

District	Cycle (days)	Daily Travel + Regular Routine Model	
		Daily μ (sd)	Regular Routine μ (sd)
Yuexiu	11	-0.95 (0.23)	-3.10 (4.04)
Tianhe	11	-1.75 (0.12)	-1.15 (3.28)
Baiyun	12	-0.32 (4.54×10^{-16})	-0.62 (2.65)
Luogang	12	-1.95 (6.32×10^{-16})	-0.41 (2.73)
Conghua	13	-1.96 (1.08×10^{-3})	-0.15 (1.86)
Zengcheng	13	-1.94 (1.36×10^{-13})	-0.42 (2.20)

References

1. McNeill, W. *Plagues and Peoples*. (Random House LLC, 2010).
2. Wesolowski, A. *et al.* Quantifying the Impact of Human Mobility on Malaria. *Science* **338,** 267–270 (2012).
3. Le Menach, A., Tatem, A. J., Cohen, J. M. & Hay, S. I. Travel risk, malaria importation and malaria transmission in Zanzibar. *Sci. Rep.* (2011).
4. Tatem, A. J. *et al.* The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malar J* (2009).
5. Stoddard, S. T. *et al.* The Role of Human Movement in the Transmission of Vector-Borne Pathogens. *PLoS Negl Trop Dis* **3,** e481 (2009).
6. Merler, S. & Ajelli, M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences* **227**, 557-565 (2010).
7. Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A. & Holmes, E. C. Phylogenetic Analysis Reveals the Global Migration of Seasonal Influenza A Viruses. *PLoS Pathog* **3,** e131 (2007).
8. Ferguson, N. M., Cummings, D., Fraser, C. & Cajka, J. C. Strategies for mitigating an influenza pandemic. *Nature* (2006).
9. Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300,** 1966–1970 (2003).
10. Lee, N. *et al.* A Major Outbreak of Severe Acute Respiratory Syndrome in Hong

Kong. *N Engl J Med* **348,** 1986–1994 (2003).

11. Pyle, G. F. The Diffusion of Cholera in the United States in the Nineteenth Century. *Geographical Analysis* **1,** 59–75 (1969).

12. Camlin, C. S. *et al.* Gender, Migration and HIV in Rural KwaZulu-Natal, South Africa. *PLoS ONE* **5,** e11539 (2010).

13. Boerma, J. T., Urassa, M., Senkoro, K., Klokke, A. & Ng weshemi, J. Z. L. Spread of HIV infection in a rural area of Tanzania. *AIDS* **13,** 1233 (1999).

14. Decosas, J., Kane, F., Anarfi, J. K., Sodji, K. & Wagner, H. U. Migration and AIDS. *The Lancet* (1995).

15. Brockmann, D. & Theis, F. Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns. *Pervasive Computing, IEEE* **7,** 28–35 (2008).

16. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782 (2008).

17. Tacconi, C., Mellone, S. & Chiari, L. Smartphone-based applications for investigating falls and mobility. 258–261 (2011).

18. Chen, J. & Bierlaire, M. Probabilistic multimodal map-matching with rich smartphone data. *Journal of Intelligent Transportation Systems* (2013).

19. Ramachandran, K. K. & Sikdar, B. Modeling Malware Propagation in Networks of Smart Cell Phones with Spatial Dynamics. *INFOCOM* (2007).

20. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE* **7,** e37027 (2012).

21. Wang, D., Pedreschi, D., Song, C., Giannotti, F. & Barabási, A.-L. Human

mobility, social ties, and link prediction. *KDD '11* 1100–1108 (2011). doi:10.1145/2020408.2020581

22. Portnoy, S. & Willson, M. F. Seed dispersal curves: Behavior of the tail of the distribution. *Evol Ecol* **7,** 25–44 (1993).

23. Vittoz, P. & Engler, R. Seed dispersal distances: a typology based on dispersal modes and plant traits. *Bot. Helv.* **117,** 109–124 (2007).

24. Clark, J. S., Silman, M., Kern, R. & Macklin, E. Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology* (1999).

25. Clark, J. S., Lewis, M. & Horvath, L. Invasion by extremes: population spread with variation in dispersal and reproduction. *The American Naturalist* (2001).

26. Tatem, A. J. *et al.* Air travel and vector-borne disease movement. *Parasitology* **139,** 1816–1830 (2012).

27. Gupta, J. K., Lin, C.-H. & Chen, Q. Risk assessment of airborne infectious diseases in aircraft cabins. *Indoor Air* **22,** 388–395 (2012).

28. Kenyon, T. A., Valway, S. E. & Ihle, W. W. Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. *... England Journal of ...* (1996).

29. Nelson, K., Marienau, K., Schembri, C. & Redd, S. Measles transmission during air travel, United States, December 1, 2008-December 31, 2011. *Travel Med Infect Dis* **11,** 81–89 (2013).

30. Raichlen, D. A. *et al.* Evidence of Levy walk foraging patterns in human hunter-gatherers. *PNAS* **111,** 728–733 (2014).

31. Condamin, S., Bénichou, O., Tejedor, V., Voituriez, R. & Klafter, J. First-passage

times in complex scale-invariant media. *Nature* **450,** 77–80 (2007).

32. Garske, T. *et al.* Travel Patterns in China. *PLoS ONE* **6,** e16364 (2011).

33. Zipf, G. K. *The P1 P2/D hypothesis: On the intercity movement of persons*. (American sociological review, 1946).

34. Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. *EPL* **81,** 48005 (2008).

35. Brockmann, D. Complex systems: Spotlight on mobility. *Nature* (2012).

36. arXiv:1010.0302v2 [cond-mat.stat-mech] 4 Nov 2010. 1–86 (2010).

37. Wen, C. H. & Koppelman, F. S. The generalized nested logit model. *Transportation Research Part B: Methodological* (2001).

38. Bhat, C. R. A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological* (1995).

39. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484,** 96–100 (2013).

40. Lessler, J. *et al.* Evidence for Antigenic Seniority in Influenza A (H3N2) Antibody Responses in Southern China. *PLoS Pathog* **8,** e1002802 (2012).

41. Lessler, J. *et al.* Location-specific patterns of exposure to recent pre-pandemic strains of influenza A in southern China. *Nat Commun* **2,** 423 (2011).

42. Google. (2011) Google Earth (Version 6) [Computer program]. Available at http://www.google.com/earth/download/ge/... (Viewed from 2009 to 2011)

Appendix

Supplementary Method

Models used in this study are as follow:

Fixed Daily Travel Models

Model 1.1 Null Model

$$Pr(Mn < d) = (F(d))^n \tag{Model 1.1}$$

Model 1.2 Models with Temporal Heterogeneity

$$\begin{aligned} Pr(Mn < d) = {} & \left(F(d|national\ holiday)\right)^{n_1} \\ & \times \left(F(d|non - holiday)\right)^{n_2} \end{aligned} \tag{Model 1.2.1}$$

$$\begin{aligned} \mathrm{Pr(Mn < d)} = {} & \left(\mathrm{F(d|national\ holiday + school\ holiday)}\right)^{\mathrm{n_1}} \\ & \times \left(\mathrm{F(d|non - holiday)}\right)^{\mathrm{n_2}} \end{aligned} \tag{Model 1.2.2}$$

$$\begin{aligned} \mathrm{Pr(Mn < d)} = {} & \left(\mathrm{F(d|school\ holiday)}\right)^{\mathrm{n_1}} \\ & \times \left(\mathrm{F(d|national\ holiday)}\right)^{\mathrm{n_2}} \\ & \times \left(\mathrm{F(d|non - holiday)}\right)^{\mathrm{n_3}} \end{aligned} \tag{Model 1.2.3}$$

$$\begin{aligned} Pr(Mn < d) = {} & \left(F(d|winter\ vacation)\right)^{n_1} \\ & \times \left(F(d|summer\ vacation)\right)^{n_2} \\ & \times \left(F(d|other\ national\ holiday)\right)^{n_3} \\ & \times \left(F(d|non - holiday)\right)^{n_4} \end{aligned} \tag{Model 1.2.4}$$

Model 1.3 Models with Spatial Heterogeneity

$$\begin{aligned} & Pr\left(M_{n|urban} < d\right) = \left(F(d|urban)\right)^n \\ & Pr\left(M_{n|suburban} < d\right) = \left(F(d|suburban)\right)^n \\ & Pr\left(M_{n|rural} < d\right) = \left(F(d|rural)\right)^n \end{aligned} \tag{Model 1.3.1}$$

$$Pr\left(M_{n|DISTRICT} < d\right) = \left(F(d|Yuexiu) \times Yuexiu + F(d|Tianhe) \times Tianhe + F(d|Baiyun) \times Baiyun + F(d|Luogang) \times Luogang + F(d|Conghua) \times Conghua + F(d|Zengcheng) \times Zengcheng\right)^n \quad \text{(Model 1.3.2)}$$

Model 1.4 Models with Household Heterogeneity

$$Pr\left(M_{n|Income} < d\right) = \left(F(d|Income1) \times Income1 + F(d|Income2) \times Income2 + F(d|Income3) \times Income3 + F(d|Income4) \times Income4 + F(d|Income5) \times Income5 + F(d|Income6) \times Income6 + F(d|Income7) \times Income7\right)^n \quad \text{(Model 1.4.1)}$$

$$Pr\left(M_{n|INCOME} < d\right) = \left(F(d|INCOME1) \times INCOME1 + F(d|INCOME2) \times INCOME2 + F(d|INCOME3) \times INCOME3\right)^n \quad \text{(Model 1.4.2)}$$

$$Pr\left(M_{n|motor} < d\right) = \left(F(d|with\ motor) \times With\ motor + F(d|without\ motor) \times Without\ motor\right)^n \quad \text{(Model 1.4.3)}$$

$$Pr\left(M_{n|Member1} < d\right) = \left(F(d|Member > 3) \times Member31 + F(d|Member \leq 3) \times Member30\right)^n \quad \text{(Model 1.4.4)}$$

$$Pr\left(M_{n|Member2} < d\right) = \left(F(d|Member > 4) \times Member41 + F(d|Member \leq 4) \times Member40\right)^n \quad \text{(Model 1.4.5)}$$

Model 1.5 Combinations

Models 1.5.1-1.5.4 are based on the sub-models with the relatively best fitting among Model 1.2-Model 1.4.

Fixed Daily Travel + Regular Routine Travel Models

Similar with fixed daily travel models, the models in this series include null model, models with spatial heterogeneity, spatial heterogeneity, and household heterogeneity. Beside of the daily travel component, we also include the units representing the regular routine travel. The basic form $(F(d))^n$ is replaced with $(F(d))^n \times (F(D))^{\lceil \frac{n}{cycle} \rceil}$.

Graphs for Fixed Regular Routine Travel

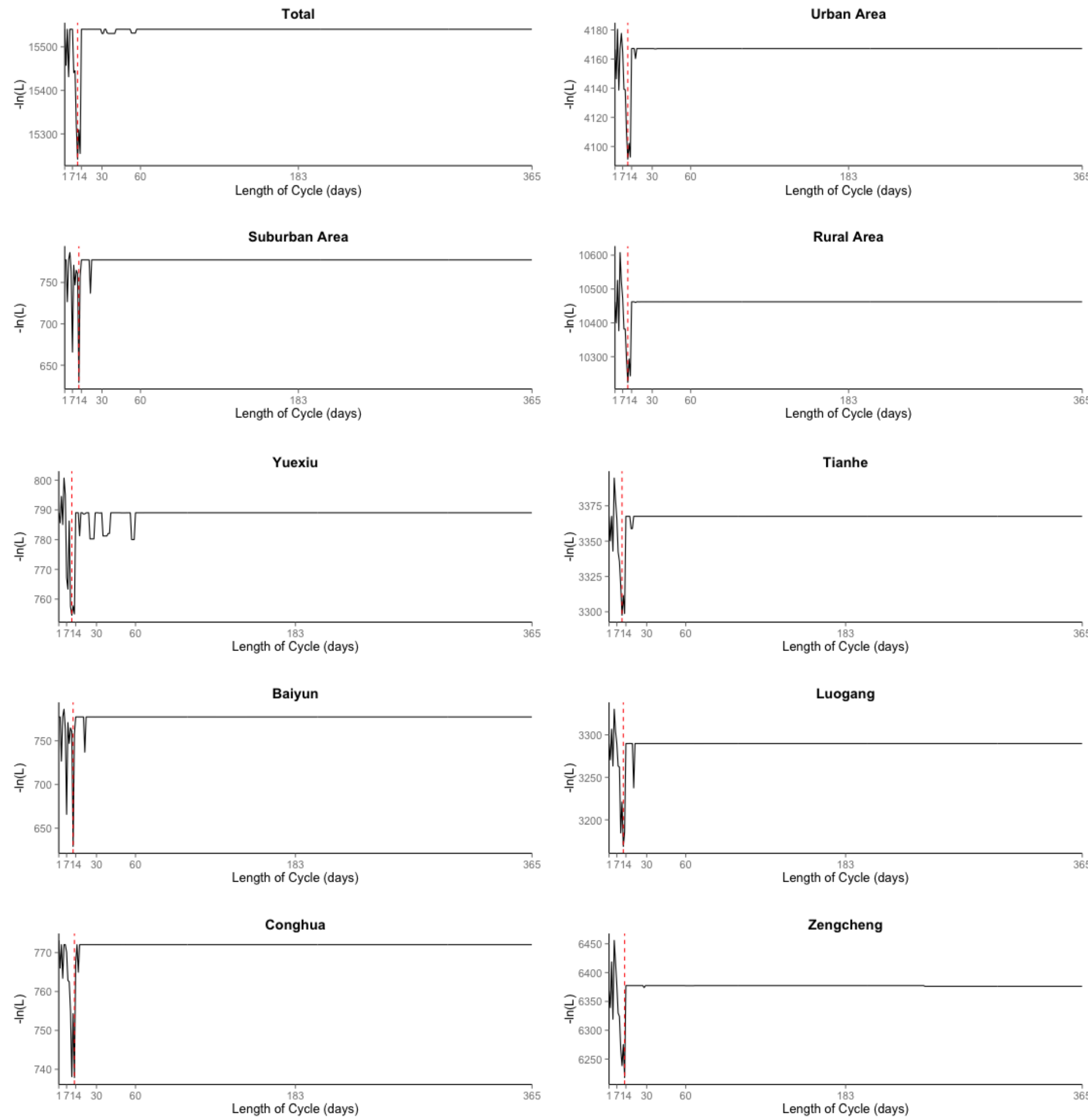

Figure 7(S). Plot of the $-ln\mathcal{L}$ value for Various Lengths of Routine Travel Cycle - 1

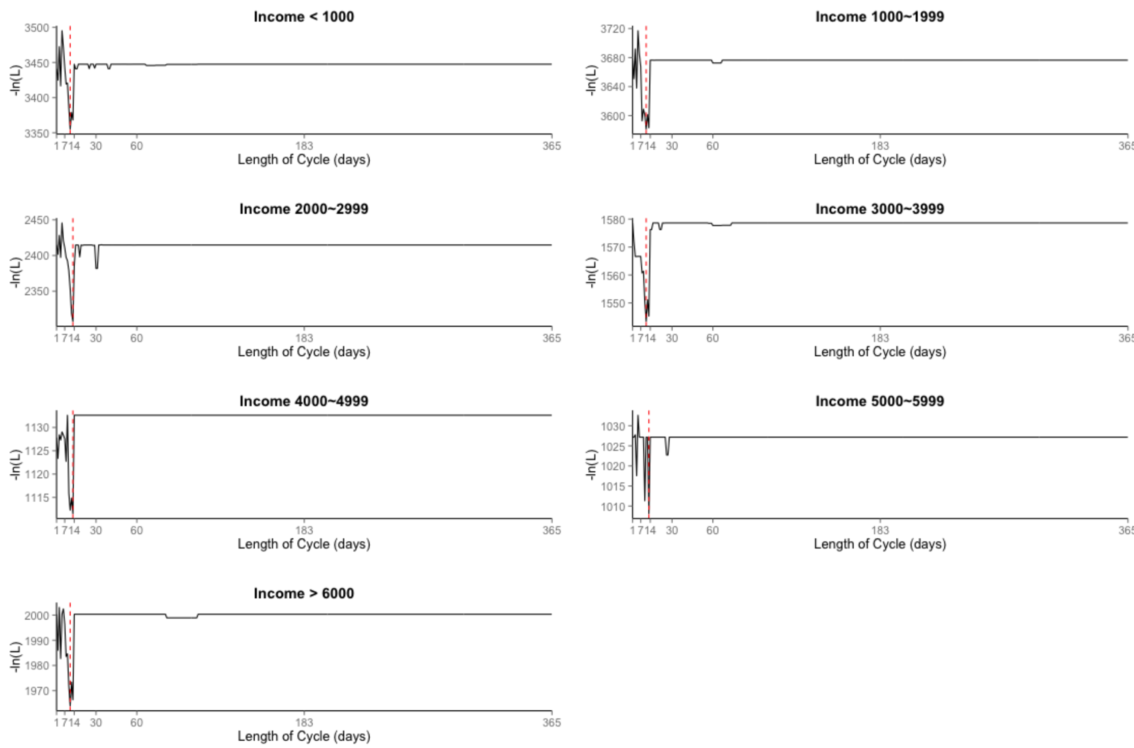

Figure 8(S). Plot of the $-ln\mathcal{L}$ value for Various Lengths of Routine Travel Cycle - 2

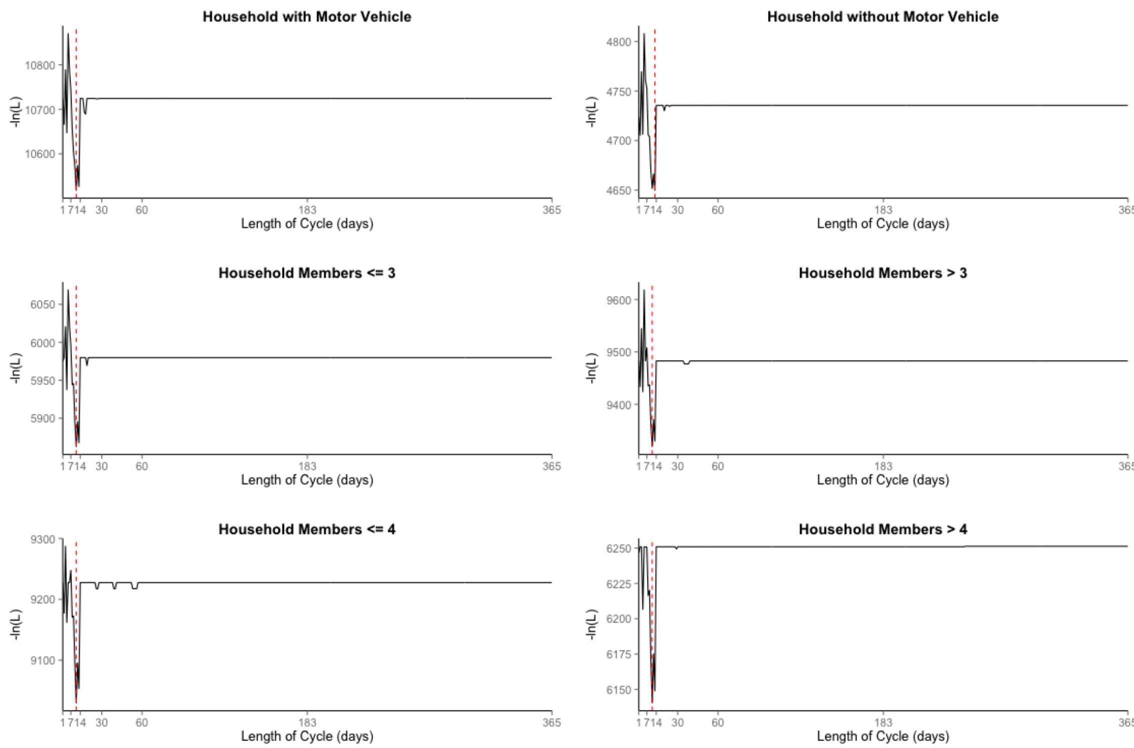

Figure 9(S). Plot of the $-ln\mathcal{L}$ value for Various Lengths of Routine Travel Cycle - 3

Curriculum Vita

Yanjie Huang

PERSONAL DATA

Phone: 301-366-1716

E-mail: yanjiezhejiang@gmail.com

EDUCATION AND TRAINING

2012-Now, ScM. Epidemiology (Infectious Disease), Johns Hopkins University

2010-2011, Intern Doctor, Peking University 9th Hospital, Beijing, China

2008-2012, B.A. Economics, Peking University

2007-2012, B.M. Medicine, Public Health, Epidemiology, Peking University

PROFESSIONAL EXPERIENCE

Student Researcher **02/2013-Now**

Project: Immune Landscapes of Human Influenza in Households, Towns and Cities in Southern China (Advisors: Derek Cummings, David Smith)

Student Researcher **01/2014-Now**

Division of Pediatrics Infectious Disease, School of Medicine (Supervisor: Pranita Tamma)

Undergraduate Student Researcher **07/2011-06/2012**

Tuberculosis (TB) Warning System Research (Advisor: Zhongwei Jia)

PUBLICATIONS

1. Zhang Long, **Huang Yanjie**, Xu Teng, Meng Xiangrui, Mao Yi, Wang Zhilong, Zhou Hong. *Study on the Current Status of Sleep Quality and Relevant Factors of Pregnant Women during Third Trimester.* Maternal and Child Health Care of China 2011; 30: 4710-4712
2. Li Wenbin, Tang Kai, Zhang Wei, Yan Wei, You Gang, Li Shaowu, Zhang Long, **Huang Yanjie**, Jiang Tao. *Relationship between Magnetic Resonance Imaging and Molecular Pathology in Patients with Glioblastoma Multiforme*. Chinese Medical Journal 2011, 124(17): 2589-2592
3. Meng Jia, Chao Wang, Yumei Zhang, Yingdong Zheng, Long Zhang, **Yanjie Huang**, Peiyu Wang. *Sugary beverage intakes and obesity prevalence among junior high school students in Beijing - a cross-sectional research on SSBs intake*. Asia Pacific Journal of Clinical Nutrition 2012, 21(3): 425
4. **Huang Y**, Carroll KC, Cosgrove SE, & Tamma PD. Determining the optimal ceftriaxone MIC to trigger ESBL confirmatory testing. (Accepted by Journal of Clinical Microbiology)