Authors' responses to the reviewers' comments on the paper

Modeling urban taxi service with e-hailings: A queuing networks approach

under full paper review for *ISTTT 23 symposium*

The authors are very grateful to the three reviewers, as well as the ISTTT23 organizers and chairs, for their constructive comments in the first-round review. We have now revised the manuscript in response to these comments. The following responses detail the changes we have made to the manuscript.

Response to Reviewer 1

- This paper presents a combined queue-theory based model with two different types of queues, taxi-passenger matching and the region-to-region traffic, respectively.
  For the taxi matching queueing model (taxi subsystem), there are two types of service, the traditional taxi and e-hailing taxi, which are assigned with different service rates. The taxi matching queueing model first distinguishes passenger and taxi queues as separate queues in a synchronized queueing system, it is then approximated using a single queue in an asymptotical fashion.
  For the region-based traffic queue model (road system), each region is simplified as a queueing system M/M/c, where c is the MFD-like regional flow capacity.
  In general, the reviewer feels that while the approach taken in this manuscript is novel, its major assumptions may not be consistent to empirical evidences, and some of the modeling efforts seem too simplified, compared with existing extensive studies on taxi matching and dynamic transportation network modeling.

  Authors' response:

  We are glad that the reviewer recognizes the novelty of our proposed approach. As stated in the manuscript, there are three main challenges in the taxi system modeling: (1) modeling the spatial heterogeneity; (2) considering network externalities, and (3) the role of stochasticity. Spatial heterogeneity has been acknowledged as an important issue and researchers have generally addressed this with region-specific models on passenger and vehicle arrivals rather than one overall model for the entire city/system. However, these existing studies do not provide any guidelines on the appropriate spatial-temporal level for modeling of taxi systems. The other two issues have not been addressed in the literature, to the best of our knowledge. Specifically, the taxi system interacts with urban road system in a way that more (or less) taxi flows will be more (or less) likely to lead congestion and increase (decrease) taxi travel time. But we have not seen any discussions on the interactions, even simple inclusions of stochastic travel time. The role of stochasticity is also critical in particular during vehicle-passenger bilateral searching. But, most studies are built upon simplistic assumptions (e.g. evenly distributed passengers and vehicles) or classical matching functions (e.g. Cobb-Douglas product function) of average performance metrics (e.g. average waiting/searching time). In general, there is a lack of accounting of the stochasticity of the arrivals in the modeling of vehicle-passenger bilateral searching.

  In this study, we address all three challenges using the novel queueing network model approach.

  - **Spatial heterogeneity.** In the queueing network, we propose a single $M/M/1$ queue for every subdivision of the taxi system to capture the dynamics of vehicle-passenger bilateral searching. Each queue has its own passenger and vehicle arrival rates, and service rates. The revised manuscript now rigorously (via extensive statistical analysis) characterizes the appropriate spatio-temporal resolution at which the Poisson assumption holds.
  - **Network externalities.** Together with the $M/M/1$ queue for the vehicle-passenger bilateral searching, we use a $M/M/c$ queue for each subdivision in order to capture vehicle movements over the urban road network. The connections between $M/M/c$ and $M/M/1$ queues include any delays on urban road network into taxi system dynamics.

- **Role of stochasticity.** While the $M/M/1$ queue makes specific assumptions of Poisson arrivals and First-Come-First-Serve [FCFS] service discipline, the method can yield not only an average system performance measurement, but also closed form expressions for metrics of interest. For instance, the waiting time distribution or the distribution of the number of waiting taxis and/or passengers. The output provides us a holistic view over not only the taxi dynamics during vehicle-passenger bilateral searching, but also, more importantly, interactions between the taxi system and the urban road network. It has been noted in the literature that it is crucial to understand the interactions between the taxi system and the urban road networks.

  Moreover, the empirical studies also highlight the advantages of queueing theoretic approaches over, for instance, the Cobb-Douglas production functions. Observe that the inputs to the queueing network (the arrival and service rates, and the routing probabilities) are (mostly) observable and can be estimated using statistical methods. However, the parameters in the Cobb-Douglas production functions are hard to calibrate with real data, and it is in general difficult to interpret these parameters in light of the available data.

We perform various statistical tests to obtain the appropriate spatio-temporal resolution justifying the Poisson assumption.

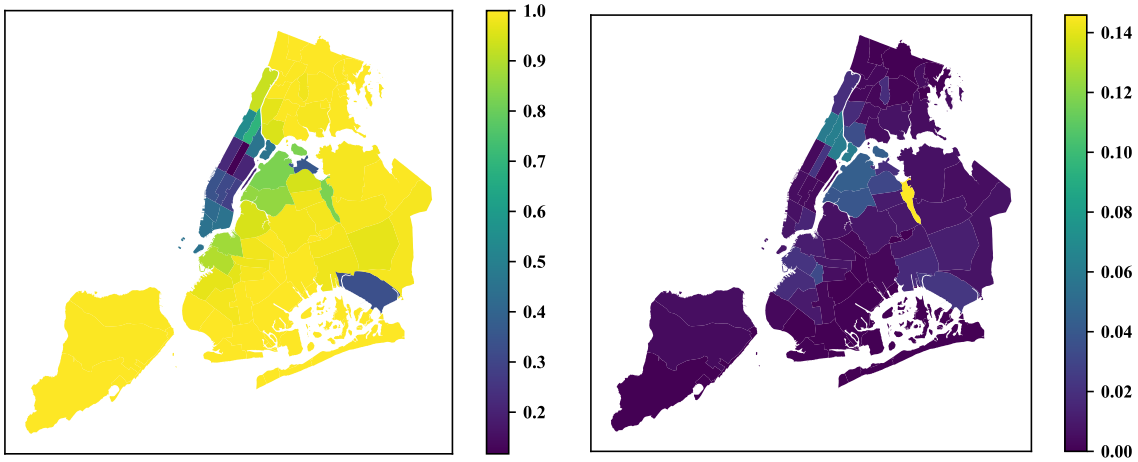
- We have now performed extensive hypothesis testing of the Poisson null assumption, and demonstrate statistically significant support at the 95% confidence level for our proposed spatial scale of 71 communities in New York City (NYC). However, we slightly adjust the temporal scale from 5-min to 1-min interval to make sure almost all major zones (i.e. generally with high ridership located close to downtown areas) do not reject the null hypothesis of Poisson arrivals at the confidence level of 95%. The key findings from empirical evidences are summarized in the following responses, as well as in the revised manuscript.
- Compared to the existing studies on taxi matching (or bilateral searching)[1, 2], the proposed queueing theoretic approach (using the synchronized $SM/M/1$ queue) is simple, but demonstrates some superior qualities over traditional approaches in the transportation engineering literature. As noted before, the widely used Cobb-Douglas product function only provides an aggregated average performance measurement. More importantly, these complicated methods are in general hard to calibrate in practice; for instance, regional effects, vacant taxi parameter, and the demand parameter in the Cobb-Douglas meeting rate function. On the other hand, the queueing theoretic approaches do not face these parameter calibration issues. Furthermore, it is our belief (based on our statistical analysis) that this model can explain more of the stochastic details in the vehicle-passenger bilateral searching. In particular, the high volume taxi data makes it possible to determine appropriate spatio-temporal scales with Poisson arrivals, which address the concerns over Poisson assumptions in the $M/M/1$ queue.
- Although the queueing network model is built upon one homogeneous time period (e.g. 1 or 2 evening peak hours) in this study, it can be easily extended to a longer time period even with heterogeneous taxi behaviors among hours. One possible solution is the pointwise stationary approximation to include the non-homogeneous taxi flows. However, the extension is out of the scope for this study thus we do not consider that in detail here.

- 1. Assumptions on fixed ratios of mode splitting. The proposed model for taxi contains two modes, traditional taxi TTS and e-hailing taxi ATS. It is assumed by the authors that the mode split can be simply modeled by a fixed probability (as "The probability of passengers at taxi subsystem m choosing ATS", and "The probability of passengers at taxi subsystem m choosing TTS"). There is little discussion on such probabilities, which seems to be arbitrary. The reviewer believes that such probabilities, if the authors still prefer to use the simple model splitting, should play critical role on the service mode selection for the passengers, and merits more discussion for this study.

Authors' response:

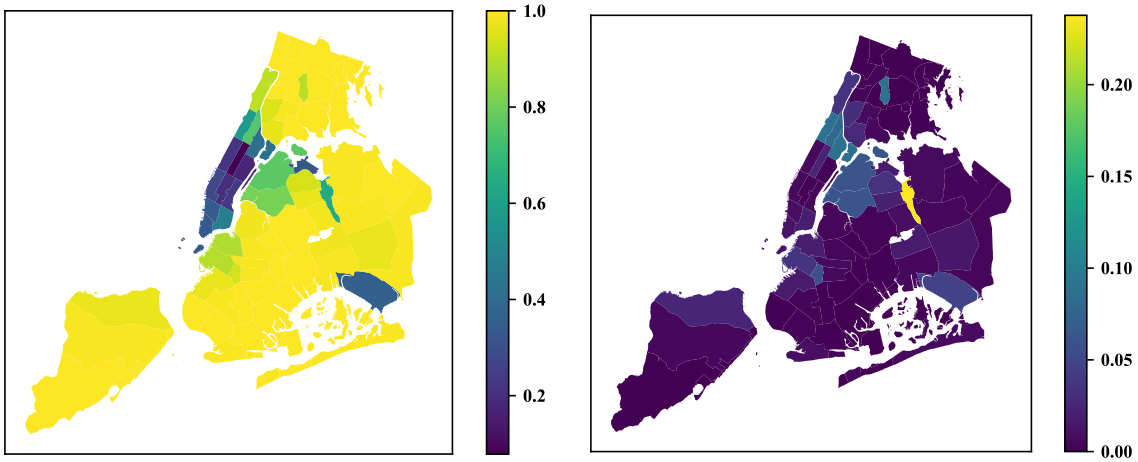
Using fixed modal split probabilities in this study depends on the spatio-temporal resolution of the study area. As the reviewer intuited, if the study focuses on all hours during the day, including both peak and off peak, the modal split *cannot* be represented by a fixed value due to many influencing factors. However, here we develop two separate case-studies, including one hour peak and another one hour off-peak. This reduces variance of the modal split probabilities. Furthermore, we aggregate passenger pickups at 71 community districts and 1 minute interval (empirical evidence for such aggregation is shown in response to comments 3 and 4).

Following the experiment settings, we summarize the probability of choosing ATS in Fig 1.



(a) Mean in peak hour

(b) Variance in peak hour



(c) Mean in off-peak hour

(d) Variance in off-peak hour

Figure 1: Probability of passengers choosing ATS in every minute and spatial unit

Specifically, Fig. 1 (a) reveals much higher market share of ATS in outer Manhattan, which is in line with official statistics. Fig. 1 (b) presents the estimated variance of the modal split probability across spatial units. It is apparent that most spatial units are within very small variance of less than 0.04, and with a small variance-to-mean ratio of less than 0.05. These statistics clearly support the assumption of fixed modal split probabilities **during peak hours**.

We summarize statistics for off-peak hours and observe similar findings, in Fig. 1 (c) and (d).

- 2. Assumptions on identical modeling for TTS and ATS and their service rates. The authors model these two types of modes in the same framework. Both types of service are modeled as a M/M/1 queue, with the only difference as the service rates. These two types of service may have multiple distinguishable differences in term of user availability, response towards ride request, cruising behavior and earning targets, which all may affect the performance and therefore the proper modeling structure. Even these might be ignored and a highly simplified M/M/1 queueing model is applied as is in this paper, it is still questionable how the authors retrieve the service rates for these two types of service in the case study. The authors use the segmented shortest path travel time as the service time for the TTS, which is different from that of the ATS. Such results seem to be a gross approximation of the service time/rate for the TTS, as many noises may not be ruled out, such as congestion effects, regional aggregation effects, etc., which is also related to the assumption on homogeneous regions as mentioned in this review later.

Authors' response:

Before describing the service rate measurements, we clarify two key points. First, the $M/M/1$ queues for both types of taxi services are considered to model the zonal level, other than taxi stands or points of interest. Thus, we segment empty taxi trips into different zones if necessary. Second, the vehicle passenger-searching time only contains two types of segments, as shown in Fig. 2, including empty trips fully inside one region (e.g. t_1 and t_2) and empty trips partially

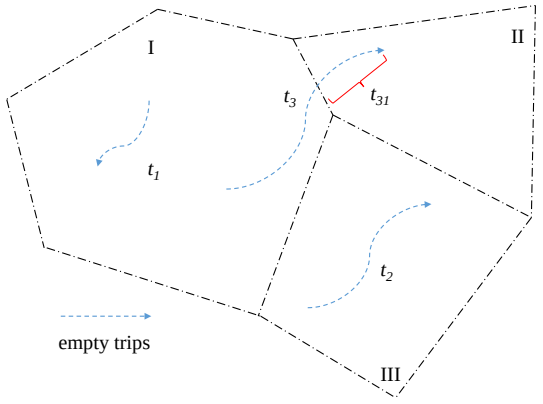


Figure 2: Illustration of service rate

inside one region (e.g. t_{31} other than t_3). The measurement can be implemented directly upon Uber trajectory data. However, we propose segmented shortest path travel time to simulate empty TTS trips, due to data availability. As the reviewer notes, this may introduce noise - such as congestion and regional aggregation effects. We believe our study reduces congestion effects by limiting the study to one hour "homogeneous" periods, and since the one hour has similar congestion effects on most passenger searching trips. We do admit that the service rate measurements depend on the spatial aggregation and assume independence among multiple sequential segments, including those segments that are for one vehicle. This is a simplification that we cannot avoid.

Coming back to the $M/M/1$ queue, we have now corrected our methodology for inferring the service rate from data. The observed vehicle passenger-searching time is obviously the sojourn time of one taxi queue (from vehicle becoming available to pickup), rather than service time as was previously stated in the original manuscript. At the regional level, it is a little hard to define a variable of matching rate between passengers and empty vehicles. However, under the $M/M/1$

modeling assumption, we can start with the sojourn time distribution to estimate service rate. Since it is well known that the mean sojourn time of the $M/M/1$ queue depends on the arrival and service rates, as shown in equations (4) and (5) of the revised manuscript. These responses are also presented at the end of Section 2.2, 'Passenger-Vehicle Matching.'

- 3. Assumptions on homogeneous regions. The modeling part and the case study both assume that each region (for traffic and taxi, both) is homogeneous. It seems that the referred homogeneity covers all the parameters, including service rates, arrival rates, etc. It is fine to do so in the modeling part; however, in the case study, there is a lack of discussion why such 71 regions can be treated as homogeneous regions expect the only claim "There are 71 community districts in total that are connected by the urban road system." It seems that the authors did not distinguish the regions based on parameter homogeneity, but rather solely based on the existing (political/residential) districts.

  Authors' response:

  In this study, the homogeneous regions means that both the traditional street-hailing and emerging app-based taxi services have Poisson passenger and vehicle arrivals (newly online by ATS driver partners or a new shift of TTS drivers). We performed extensive statistical hypothesis testing of the Poisson null hypothesis at different spatial scales, as well as temporal intervals and homogeneous time periods. We use the following standard statistical tests: The Kolmogorov-Smirnov (KS) test adapted for discrete distributions[3] that examines whether the passenger or vehicle arrivals can be assumed to be Poisson distributed over a fixed time horizon. In addition, three additional χ^2 distribution based tests[4] (i.e., the Anscombe, likelihood, and conditional tests) are applied to test whether the arrivals satsify the Poisson null hypothesis, and more importantly, whether they are from a single homogeneous Poisson distribution. The settings for hypothesis testings are as follows:

  - The potential spatial scales are mainly based on four administrative divisions in NYC, including Borough[1], community districts[2], zip code tabulation area [ZCTA] [3], and census tract [4]. Note that we use administrative divisions rather than grid based spatial scale, since most socioeconomic variables are only available at administrative divisions and modeling at administrative divisions will be much easier to measure socioeconomic impacts on taxi activities in future studies.
  - We test seven different count intervals, that are 1-min, 5-min, 10-min, 15-min, 20-min, 30-min, and 60-min. In other words, we count TTS, ATS, or both TTS and TTS pickups (or vehicle arrivals) every count interval then tests whether the flow can be described using a Poisson process.
  - In addition, we also test the selection of the homogeneous time period, considering time-of-day and day-of-week effects. Regarding the peak (or off peak) hour, we include three difference cases, including 1-hour period (peak: 6pm to 7pm, or off peak: 10am to 11am), 2-hour period (peak: 5pm to 7pm, or off peak: 9am to 11am), and 3-hour period (peak: 5pm to 8pm, or off peak: 9am to 12pm). Moreover, we classify the weekdays from Mondays to Thursdays, compared to all seven days case.

  Overall, regarding the spatio-temporal scale, we can still use the proposed spatial divisions of 71 community districts but should slightly decrease the 5-min count interval to 1-min. The empirical

[1] 5 in total and each is with an average area of 60.4 mi^2, see `https://en.wikipedia.org/wiki/Boroughs_of_New_York_City`

[2] 71 in total if we include regions of airports and parks, and each is with an average of 4.3mi^2, see `https://www1.nyc.gov/site/planning/community/community-portal.page`

[3] 214 in total, and each is with an average of 1.4 mi^2, see `https://www.census.gov/geo/reference/zctas.html`

[4] 2165 in total, and each is with an average of 0.14 mi^2, see `https://www.census.gov/geo/reference/gtc/gtc_ct.html`

findings are summarized as follows: (Note that we only present the hypothesis test results for passenger pickups and vehicle arrivals during off peak hours, since the variations in hypothesis test results are similar between off peak and peak hours.)

- Fig. 3 to 10 show the percentage of zones *not* rejecting the Poisson null hypothesis using the 4 methods, 7 count intervals, and day of the week. It is apparent that the smaller count interval generally leads to more zones not rejecting the Poisson hypothesis, across the plots. Although the hypothesis tests using the corrected-KS reveals similar percentages from 1-min to 60-min count intervals, the homogeneous Poisson tests generally reject the null hypothesis that the arrival counts are from a single homogeneous Poisson distribution if we have a larger count interval. Thus, we reduce the count interval from 5-min to 1-min in the revised manuscript.
- Fig. 3 to Fig. 6 summarize the percentages of zones where passenger pickups can be

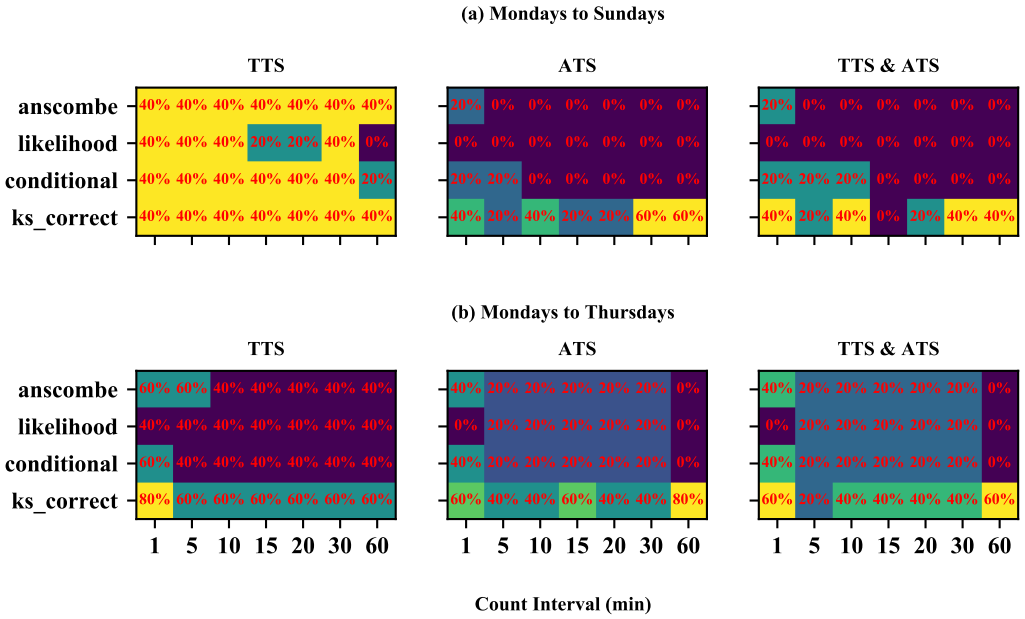


Figure 3: Hypothesis test results for passenger pickups at Boroughs in one-hour off peak

assumed as Poisson distribution at four levels of spatial scale, respectively. Within our expectations, the too large or small aggregation will lead to much fewer spatial units not rejecting hypothesis testing, since large zones have more spatial interactions and heterogeneity and small zones usually do not have any rides. Both community district and ZCTA aggregation have higher percentages. And community district aggregation generally has a slightly higher percentage, regardless of methods, taxi services, count interval and day of the week.

- Fig. 7 to Fig. 10 exhibit the percentages of zones where vehicle arrivals (newly online by ATS driver partners or a new shift of TTS driver) can be assumed as Poisson distribution at four levels of spatial scale, respectively. Similar as test results for passenger pickups, both community district and ZCTA aggregation reveals higher percentages. In particular, the percentages resulted from ATS vehicle arrivals, are sometimes close to 100%.
- To sum up, we choose community district aggregation. Except for the hypothesis testing for Poisson assumption, we also consider the homogeneous regions for road queues. In general, such regions are relatively larger than those for taxi system. To balance the both systems

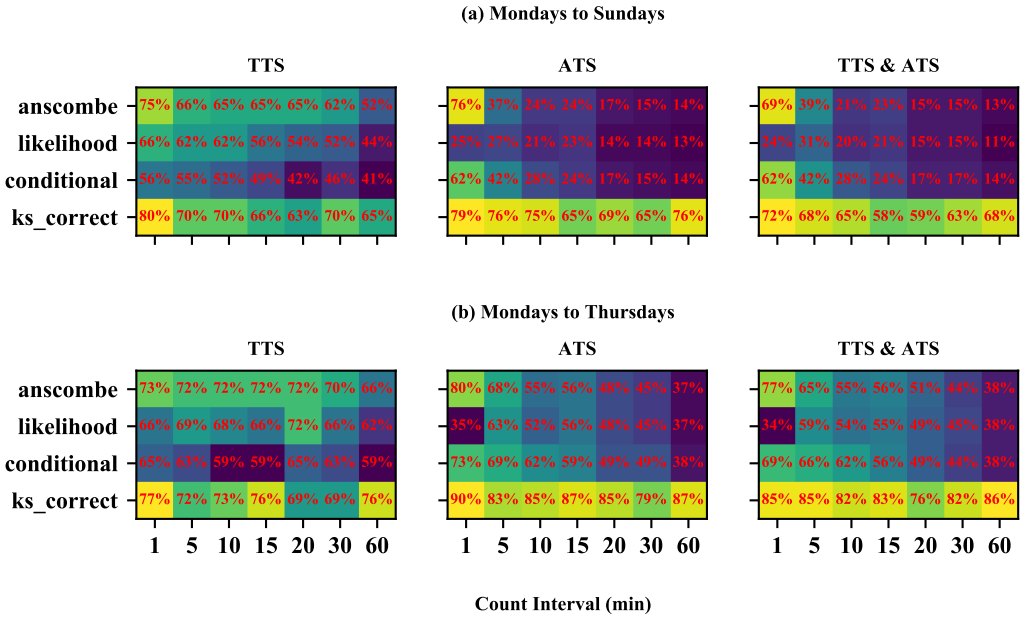


Figure 4: Hypothesis test results for passenger pickups at Community Districts in one-hour off peak

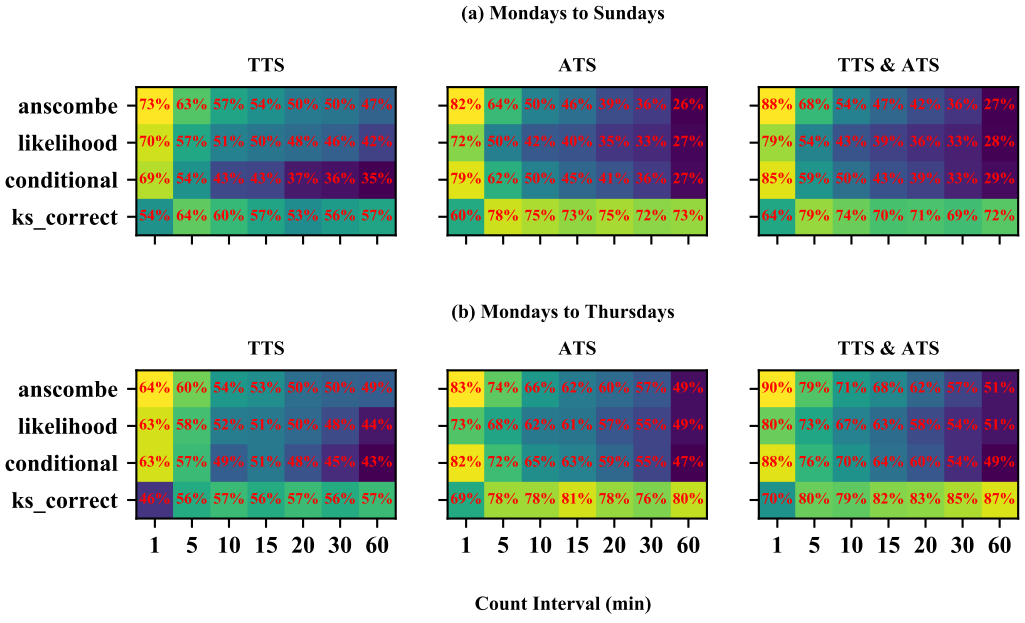


Figure 5: Hypothesis test results for passenger pickups at ZCTA in one-hour off peak

and set up same homogeneous regions, the community districts are also appropriate spatial scale.

Regarding the homogeneous period selection, we would like to shorten our study period to weekdays and one-hour peak (or off peak), mainly depending on the following empirical evidences:

- Fig. 4, 11, and 12 compare the hypothesis results for community district aggregation of

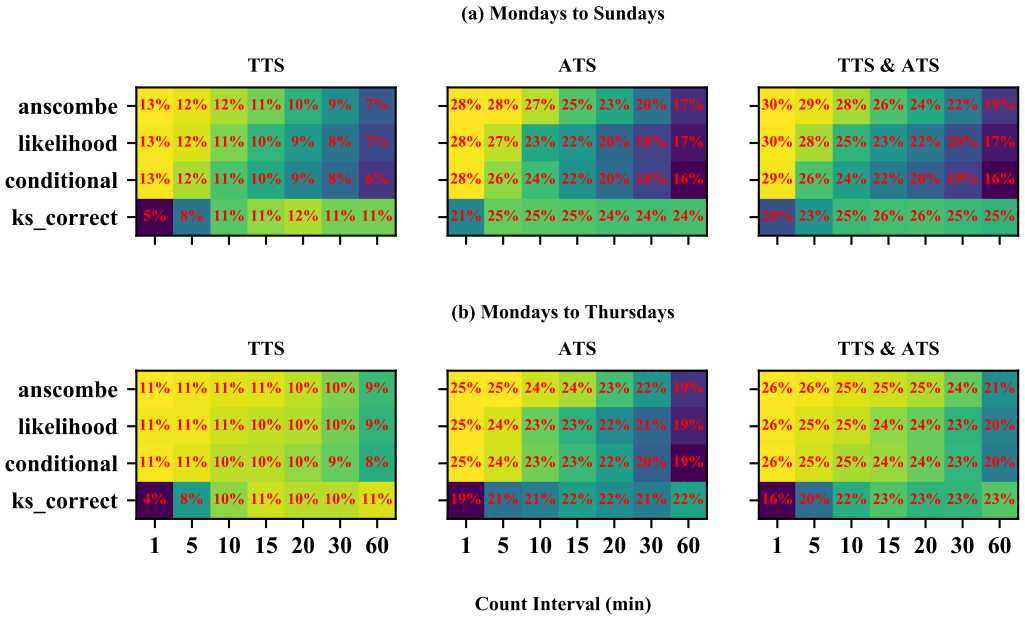


Figure 6: Hypothesis test results for passenger pickups at Census Tracts in one-hour off peak

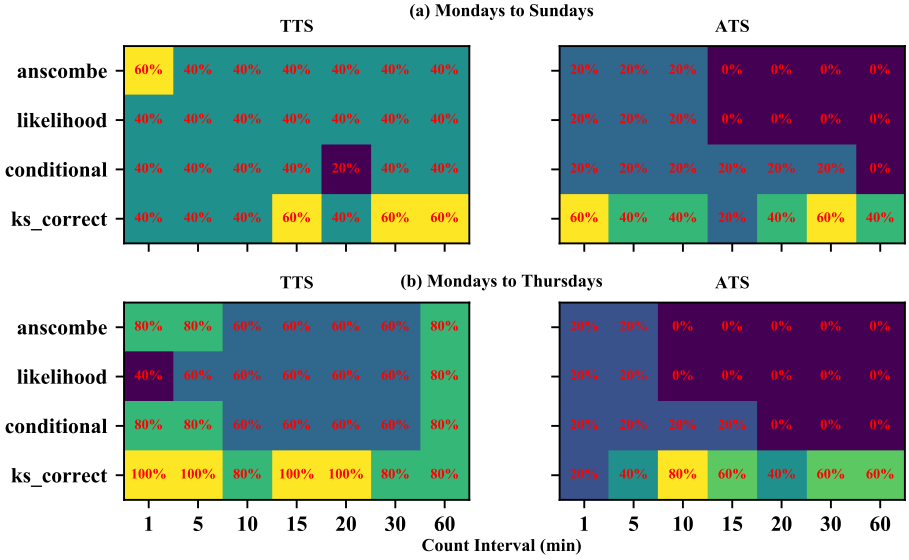


Figure 7: Hypothesis test results for vehicle arrivals at Boroughs in one-hour off peak

passenger pickups across different levels of off peak hours, as well as day of the week. As number of hours included into off peak increase, fewer community districts are not rejecting Poisson distribution. In addition, limiting to the weekdays from Monday to Thursday can slightly increase percentages of significant zones.

- Fig. 8, 13, and 14 compare the hypothesis results for community district aggregation of vehicle arrivals across different levels of off peak hours, as well as day of the week. There are no big differences in the percentages by week of the day, as well as number of hours in off peak period. However, introducing more hours or focusing on weekdays can lead to very

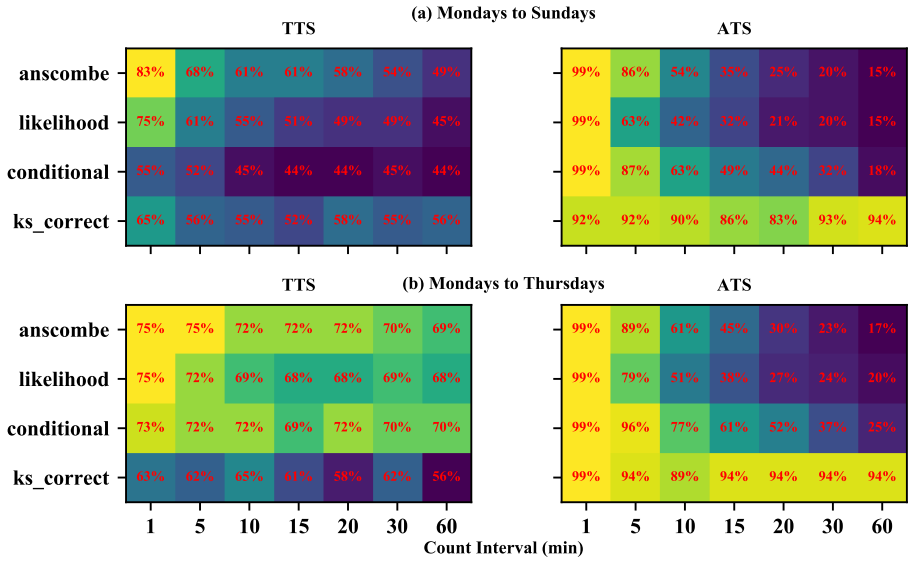


Figure 8: Hypothesis test results for vehicle arrivals at Community Districts in one-hour off peak

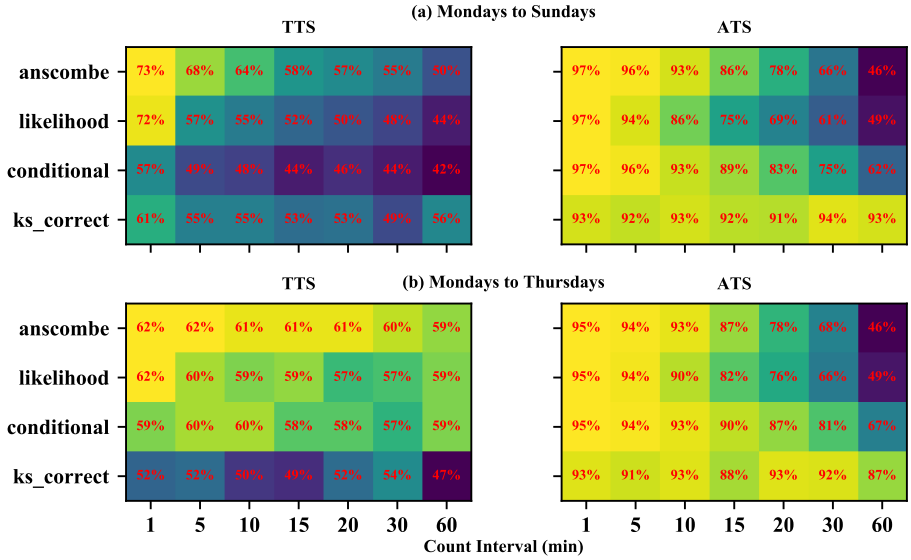


Figure 9: Hypothesis test results for vehicle arrivals at ZCTA in one-hour off peak

small increases in percentages.

- 4. Assumptions on Poisson arrivals. Empirical results in Figure 8 clearly show that most cases the passenger and taxi arrivals do not follow the Poisson distribution. The so-called fitted Poisson, may not be a good candidate distribution to capture realistic arrival processes, except the Uber vehicle arrivals for peak hours, and the taxicab passengers arrivals for off-peak hours.

  Authors' response:

  As stated in our responses to your previous comment, we utilize community district aggregation and 1-min count interval in the revised manuscript. Meanwhile, we limit our case study to 1 hour

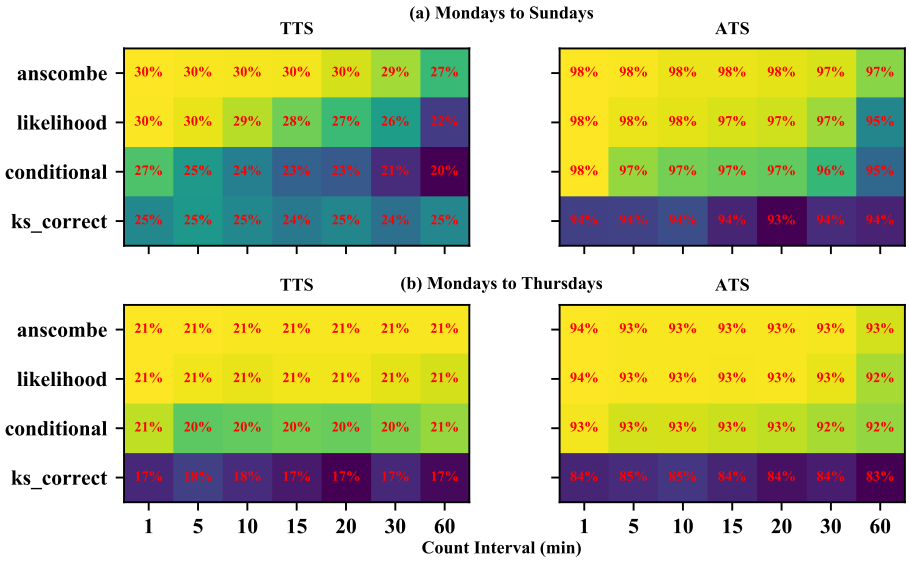


Figure 10: Hypothesis test results for vehicle arrivals at Census Tracts in one-hour off peak

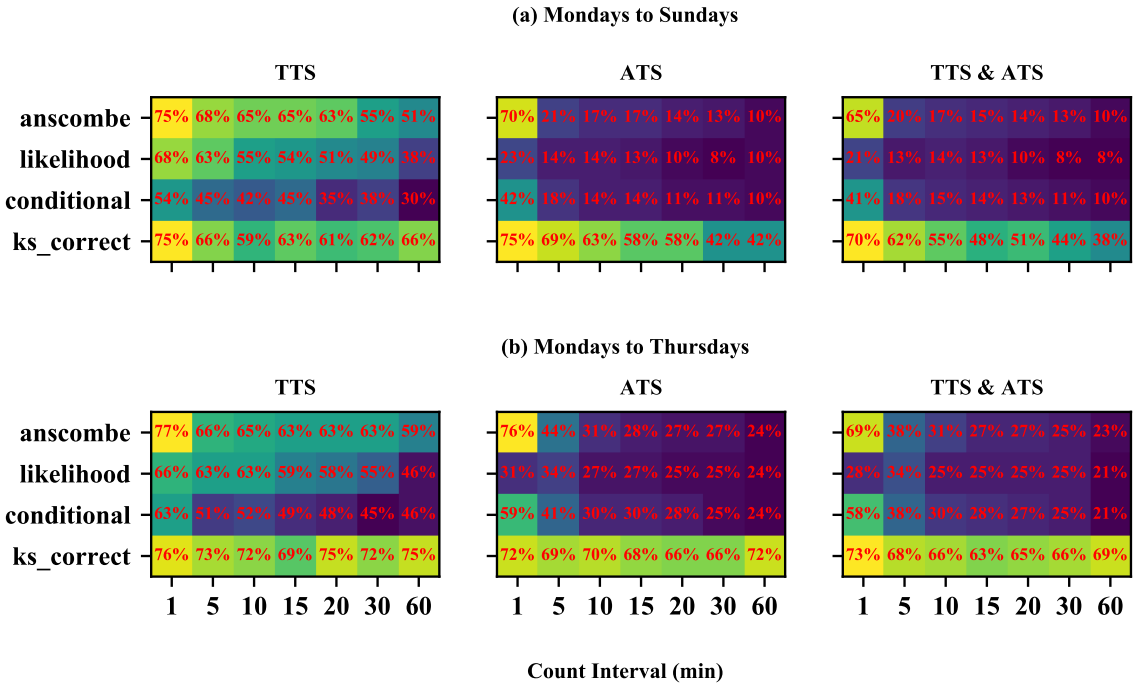


Figure 11: Hypothesis test results for passenger pickups at Community Districts in 2-hour off peak

peak (or off peak) period in only weekdays from Monday to Thursday. With these adjustments, we find the both passenger and vehicle arrivals can be assumed to follow a Poisson process, in more than 75% of community districts, as shown in Fig. 4 and 8. In addition, we plot the spatial distribution of community districts not rejecting the Poisson distribution in Fig. 15 and 16. Both figures indicate that those community districts, rejecting the Poisson assumption, are primarily located in remote suburban areas, generally with rare TTS and ATS activities. In

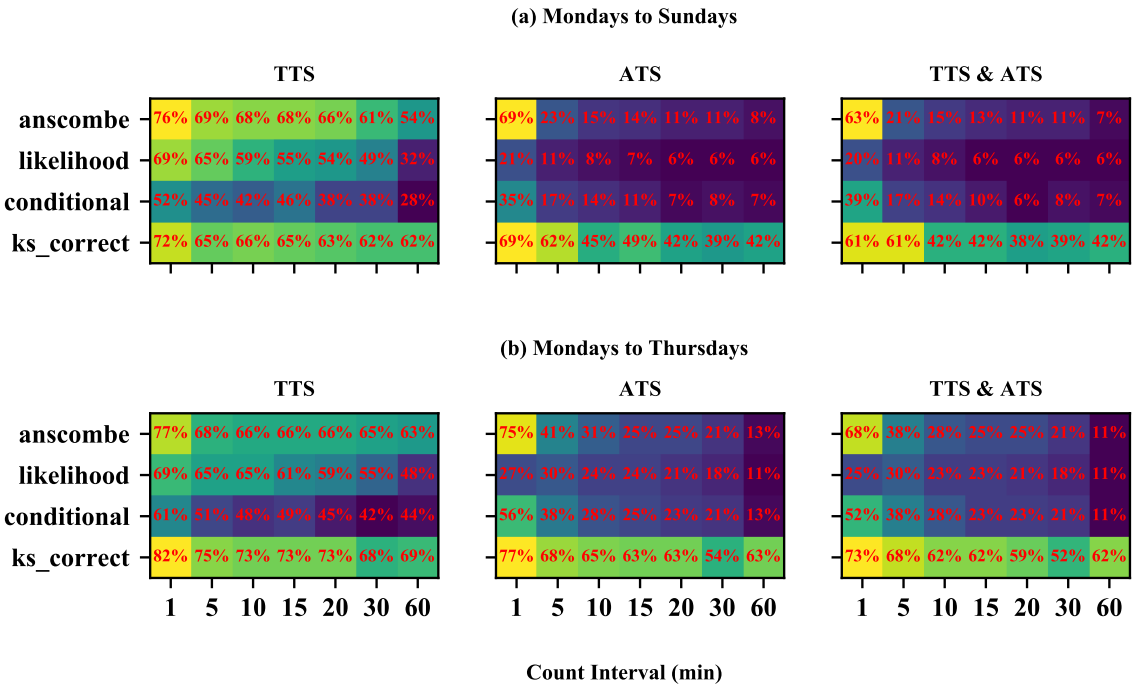


Figure 12: Hypothesis test results for passenger pickups at Community Districts in 3-hour off peak

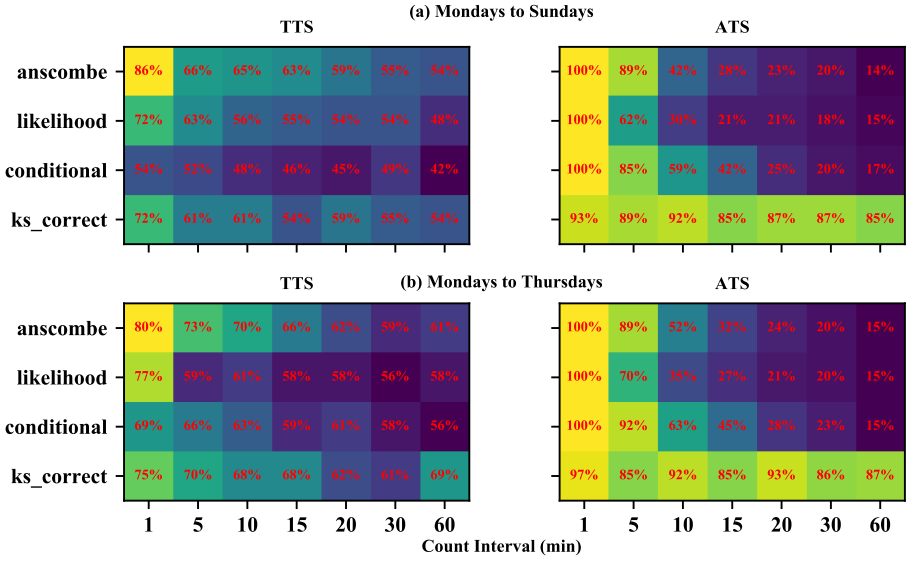


Figure 13: Hypothesis test results for vehicle arrivals at Community Districts in 2-hour off peak

NYC, most taxi activities concentrate in downtown and midtown Manhattan, as well as the two airports, downtown Brooklyn, and downtown Queens. Our hypothesis tests strongly support the Poisson process assumption on passenger and vehicle arrivals in those areas. For more details on TTS and ATS activities and facts in NYC, you can refer to 2018 NYC TAXI FACT BOOK (`http://www.nyc.gov/html/tlc/downloads/pdf/2018_tlc_factbook.pdf`).

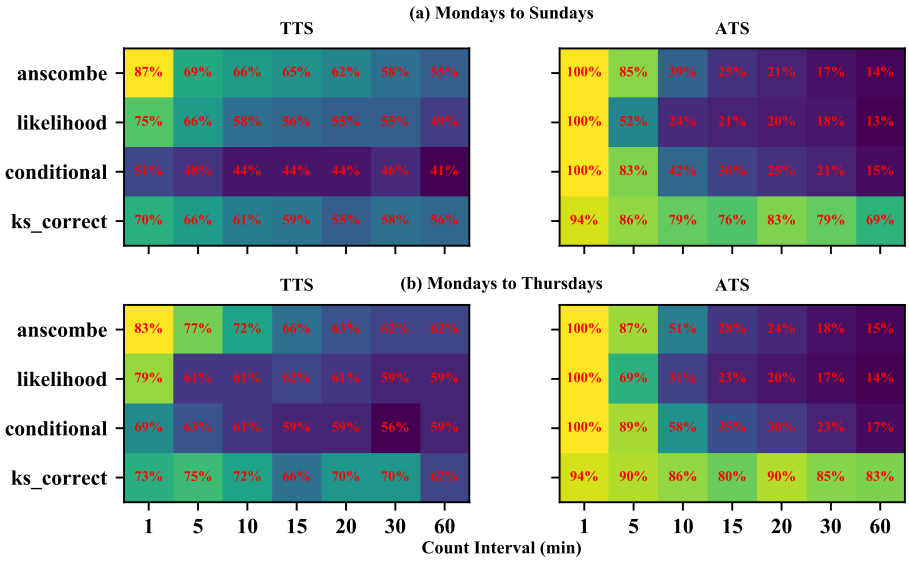
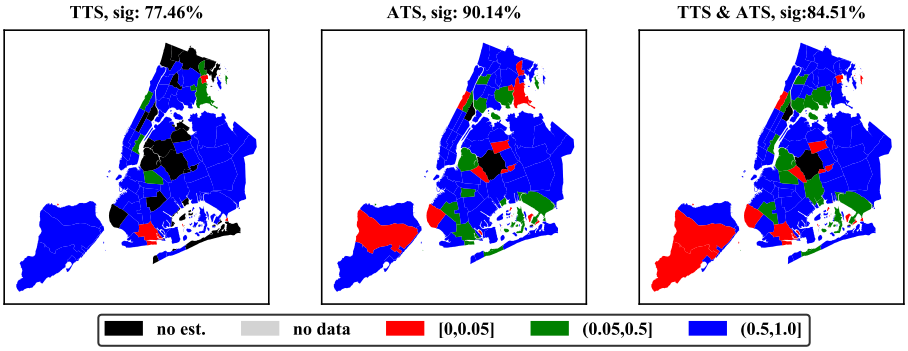
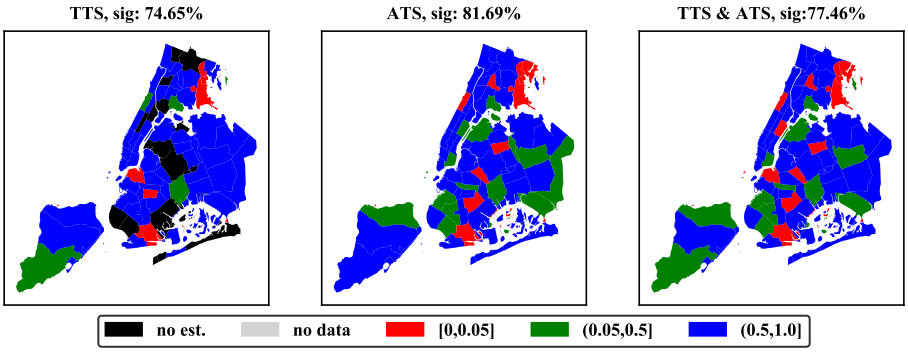


Figure 14: Hypothesis test results for vehicle arrivals at Community Districts in 3-hour off peak

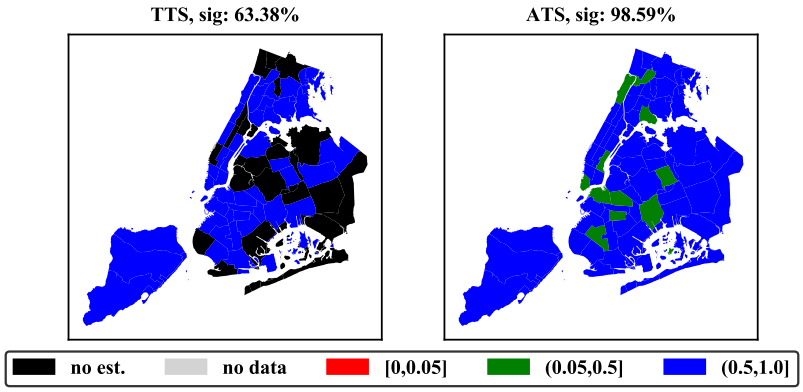


(a) 1-min count interval in 1-hour off peak

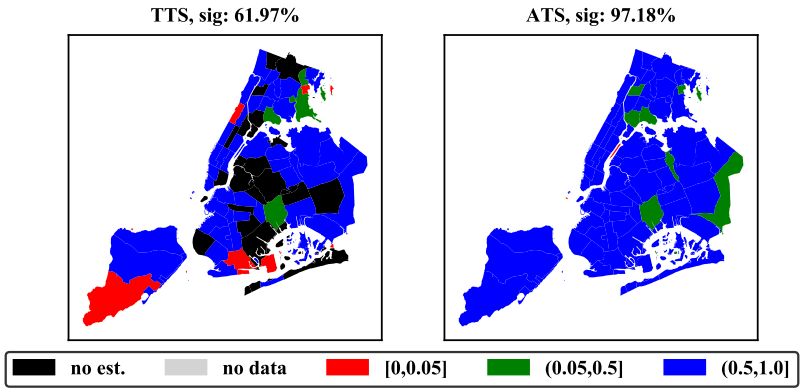


(b) 1-min count interval in 1-hour peak

Figure 15: Hypothesis test results for passenger pickups by Community Districts in weekdays. Note: 'sig' indicates percentage of community districts not rejecting Poisson distribution, represented by blue and green color



(a) 1-min count interval in 1-hour off peak



(b) 1-min count interval in 1-hour peak

Figure 16: Hypothesis test results for vehicle arrivals by Community Districts in weekdays. Note: 'sig' indicates percentage of community districts not rejecting Poisson distribution, represented by blue and green color

- 5. Assumptions on service rates with state-dependence. In Section 2.4, the authors propose eq(4), which claims that if there are more vehicles than un-served riders, the service rate increases, while if the opposite happens, the service rate decreases. Such claim lacks theoretical analysis or empirical support. It is also possible that in either case, the service rate would drop from its maximum of μ_0. If the authors can provide more convincing analysis or referred literature, it is easier for the reader to understand the rational behind this assumption.

  Authors' response:

  In the revised manuscript, we have changed the measurement methods for the service rate, as well as the definitions, at the end of section 2.2 'Passenger-Vehicle Matching.' The service rate at the zonal level is not directly observable and should be an intrinsic characteristic related to region configuration and taxi service type. However, following our response to comment 2, under the $M/M/1$ modeling assumption, we can use the sojourn time distribution and expected (or mean) sojourn time to estimate the service rate. Note that the expected sojourn time of $M/M/1$ depends on arrival rate and service rate, as shown in equations (4) and (5) in the revised manuscript. In the revised manuscript, we replace the section 2.4 with these corrections.

- 6. Assumption on asymptotical approximation. Such an approximation is one of the key assumptions made in this paper, but unfortunately it is not consistent with the modeling context.

The assumption requires the system has a long time to be stable (t goes to infinity, as asymptotically). But in the setup for this problem, one has to consider a short time period when all related parameters can be treated as constants. In the case study, such a short time period is 5 minutes. A Poisson process in such an asymptotic sense is far from the actual Poisson process, which is also suggested from Figure 8 as an example.

Authors' response:

First, we should clarify the term t. It is not the count time interval, 5 minutes, as you mentioned. It actually relates to the *relaxation time*, required for the distribution of the Markov chain modeling the state of the queueing network to reach its stationary/invariant distribution. Relative to the inter-arrival and service times in the queueing network, if t is large enough, then the proposed $SM/M/1$ can be asymptotically approximated by standard $M/M/1$ taking the flow with minimum arrival rate as inputs.

In practice, it is hard to say whether a relaxation time is long enough, and there is a substantial literature on understanding the relaxation times of birth-and-death Markov chains that is relevant to this discussion. This related literature provides an upper bound, $t < ((1+\rho)/(1-\rho))^2$ to measure such time[5]. Based on this, in both our case study's the queueing model can be assumed to 'relax' with less than 0.95 utilization rate, and a observation period of 1 min. Most spatial units meet these requirements. Only one spatial unit in midtown Manhattan has higher utilization rate of more than 0.95 at peak hour, regardless of ATS and TTS. The spatial unit with heavy load (utilization rate is about 0.97), demands more observation periods to relax. During off peak hours, there are no spatial units with utilization greater than 0.90 for either service types.

Note that we reduce the observation period from 5 minutes to 1 minute based on extensive hypothesis testing. As shown in the response to your previous comments, a the 1 minute observation interval more spatial units do not reject the Poisson null hypothesis. This study has about 1,380 observation intervals in peak (or off peak) hour case, since the peak (or off peak) case has 1 hour of each 23 days.

- Page 4. "both the ATS and TTS do not have stations". Is it better as "neither of the ATS and TTS have stations"?

  Authors' response:

  Corrected. Thanks.

- Page 4 and beyond. It is better to briefly introduce what does S M/M/1 queue mean.

  Authors' response:

  We have added two sentences immediately after the first emergence of "$SM/M/1$" in page 4 to briefly introduce it:

  "Different from regular $M/M/1$ queue with one arrival flow, $SM/M/1$ has two independent arrival flows of both passengers and vehicles thus processes synchronized passenger-vehicle pairs that match based on arrival sequences and zero matching time. Although there exist certain differences, the $SM/M/1$ queue can be further approximated by simple $M/M/1$ taking the minimum of two arrival rates as input, which are also shown in this study."

  These sentences provide a brief summary of the synchronized $SM/M/1$ queue and related works.

- Page 4. "data science approaches.." seems to be vague. What type of data science approaches are referred? Which method is utilized and how is it contributing this study?

  Authors' response:

'Data science' methods enter our study in two ways. The primary way is the statistical hypothesis testing methodology used to measure the Poisson null hypothesis. This study primarily examines three types of datasets, including passenger arrival counts under certain time interval, vehicle arrival counts under certain time interval, and empty vehicle searching time. The former two datasets are tested with Poisson distribution, corresponding to Poisson assumption of arrival flow for queue theoretic approaches. And the last one is tested with exponential distribution, corresponding to exponential distribution of service time for queue theoretic approaches. The second way in which data science methodology enters our study is through the data and summary statistics visualizations we have now provided.

These data science methods will help answer where and how many spatial units satisfy the assumptions of the queueing theoretic model. Moreover, the hypothesis test results lead to better spatio-temporal aggregation scales. Lastly, it is useful for identifying homogeneous study areas and periods. In the revised manuscript, the second-to-last paragraph is extended with this discussion.

- Page 7. 'models the how quickly...' delete 'the'.

  Authors' response:

  Corrected. Thanks.

- Page 4. Eq(1) and (2). the summation of lambda and E, O mixes a rate and two numbers of vehicles together. Should they share the same unit, say rate plus rate, or number plus number?

  Authors' response:

  Yes, they share exactly same units. The E and O in original manuscript refer to arrival rate of vehicles originating from neighboring spatial units but searching and picking up in the spatial unit. They are arrival rates as λ for newly joined vehicle rates.

  In revised manuscript, we simplified several expressions and mathematical forms with fewer symbols. For example, we replace E and O with one unified symbol of $F_{i,in}$ at spatial unit i and replace balking probability $J_i^{T,ATS}$ with probability of successfully picking up passengers $p_i^{P,ATS}$. Same replacements are applied for TTS. The new mathematical forms are simpler but easier to understand.

- Page 8. 'split the departure flow'. It is unclear without detailed splitting manner on how to exactly split the departure flow. Would the destination of trips be considered or not?

  Authors' response:

  First, we admit that the word 'split' here is a little bit confusing. To clarify, this is the routing of the departure flow to other spatial units, which depends on vehicle status and service types. Different types of vehicles are assigned with special routing probabilities for distribution over the road network. Identification of vehicle types is based on their status in incoming flows to spatial units. These clarifications are also included in the section of 'inclusions of road network performance'.

  On the other hand, we did consider the destination of trips. This is included in the routing probability matrix in two ways. First, the diagonal elements reveal the destination information for those drivers who pickup and drop-off in one same spatial unit. Second, for those drivers who pickup and drop-off in two different spatial units, we model the movements by an indirect way of 'random walk'. Instead of modeling the probability of destination choice, we derive the transfer probability of moving from one spatial unit to the neighboring spatial units. Under the 'random walk', the destination choice is the product of transfer probability. In addition, we add one virtual spatial unit in the routing matrix, denoting vehicle exiting taxi system. To sum up, our routing matrix contains not only trip destination information but also vehicles' system-exiting behaviors.

- Page 9 and beyond. The mixed use of S_t and $S(t)$ is confusing. Are they referring the same variable?

  Authors' response:

  Yes, you are right. The both symbols refer to one same variable. The $S(t)$ has been replaced with S_t.

- Page 12. More explanation or formal proof is desired why replacing all min sets with inequalities ensure the same solution. It actually looses the constraints and may lead to other solutions.

  Authors' response:

  Yes, you are right. The original replacement looses the constraints. In revised manuscript, we have introduced four additional inequalities, as equation 19, 20, 23, and 24. It is straightforward that the inequalities (equation 17 to 24) limit the variables of interests to the point of minimum values. Thus, we have an equivalent replacement for all the min sets in the proposed formulation (equation 16).

Response to Reviewer 2

- This is an interesting and largely well written paper on a topical subject. However, I struggle with the motivation. We do of course wish to understand how street-hailing and e-hailing taxis interact with each other in congested urban networks from a policy perspective, but the paper does not seem to lead to any policy conclusions. Instead, the model is offered for control or management in the taxi market. However, given that the model makes use of spatial units (also referred to as road subsystems or regions, but could be called simply zones) with no road network, I doubt whether it is suitable for control or operational management.

  Authors' response:

  We thank the reviewer for their largely positive assessment, and we attempt to address the concerns raised. This study is motivated from current research gaps in quantifying both traditional street-hailing and emerging app-based taxi services. As stated in the manuscript, existing studies have not well addressed three crucial concerns in the taxi system, including: (1) modeling the spatial heterogeneity; (2) considering network externalities, and (3) the role of stochasticity. Although the spatial heterogeneity has been acknowledged, the existing studies did not provide any guidelines on the appropriate spatial-temporal level for modeling of spatially and temporally heterogeneous taxi systems. The other two issues have not been addressed in the literature.

  More importantly, the interactions between taxi system dynamics and urban road congestion are not mentioned by majority of the literature. Thus, this study is developed to propose an unified and reliable modeling structure, not only quantifying taxi system performance metric but also filling in current gaps in external interactions with the urban road system.

  As mentioned in the reviewers comment, this study is designed at aggregated level, rather than individual, micro-level, or even urban road link level. Our goal is to quantify both the aggregated taxi performance (e.g. number of waiting passengers or searching drivers, waiting time distribution, etc.) and aggregated urban road performance (e.g. travel delays) in one 'homogeneous' region. With these performance metrics, it is possible to extend to work with zone-based control or management strategies. However, to comprehensively demonstrate this is out of the scope of this paper. Here, we also list several potential implications. First, we can find a better routing matrix to navigate empty vehicles and make the taxi system more efficient in the whole city. Second, we can find a better vehicle arrival rate in each zone and a better online (or shift) zone for traditional street-hailing taxicabs to improve scheduling and make them more competitive. Third, we can extend the current modeling structure with a state-dependent queue to capture

effects of dynamic pricing which can impact ridership and vehicle fleet policy. Furthermore, we also believe that our queueing network model will provide a holistic view of the urban taxi system, bridging the 'physical' road infrastructure and the 'virtual' infrastructure of the ATS and TTS.

- The taxi market is described as oligopolistic in many places, but I dont think this is an accurate description of the market given that e-hail drivers decide when and if to work. The e-hail companies do not control the supply, just the fare, and so they function like a crude market clearing mechanism ensuring the supply and demand balance. Hence surge pricing when e-hail taxis are in short supply. The advent of e-hail taxi services has considerably reduced market power in the taxi market, bringing fares closer to marginal costs (except for times where taxis are in short supply). In any case the concept of oligopoly does not play any role in this paper, so I suggesting deleting the terms monopolistic and oligopolistic.

  Authors' response:

  Thanks for your clarifications and recommendations. We agree on the point. Moreover, the terms do not play any role in this paper as mentioned, except to differentiate traditional and emerging taxi markets. Based on above considerations, we delete the terms 'monopolistic' and 'oligopolistic' and replace corresponding expressions with 'TTS-dominated' and 'competitive', respectively.

- I am not familiar with the literature reviewed, but it appears to be deep enough.

  Authors' response:

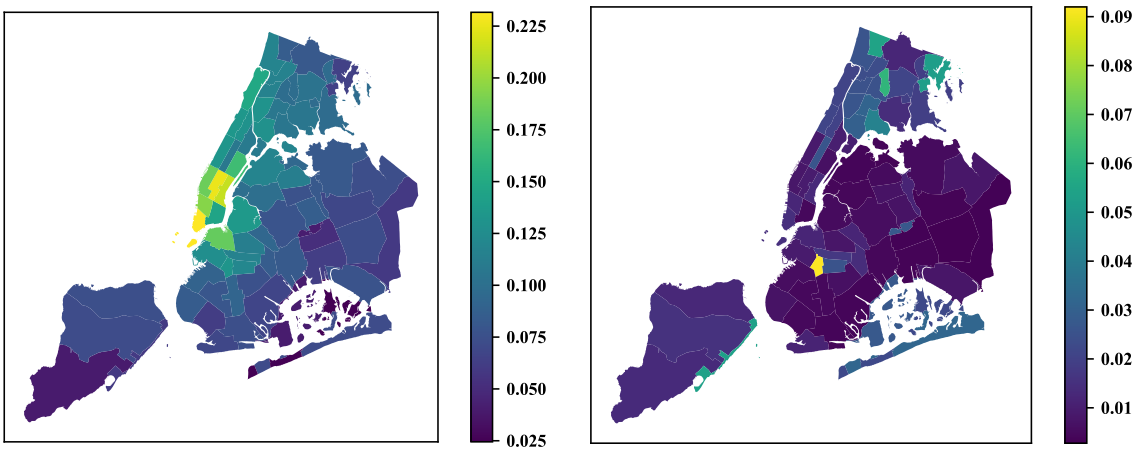
  Thank you.

- The model consists of a network of queues, specifically passengers queuing for vehicles and vehicles queuing for passengers, and allows for imbalances in passenger flows and corresponding vehicle repositioning movements. Fig. 1 neatly describes the demand and supply interactions, and Fig. presents the network of regional queues. The passenger-vehicle matching process is described (Are the Cobb-Douglas equations the well-known Cobb-Douglas production functions, and if so, where to they fit in here?). The phomenon of balking is described, but surely this is vehicle repositioning and is motivated by the prospect of a sooner or bigger job in another region? Can this be modelled accurately by a fixed probability?

  Authors' response:

  This study proposes a different approach based on queue theoretic method, rather than classical Cobb-Douglas production function. We mention the Cobb-Douglas production function in this study primarily to exhibit the current gaps in the literature and to demonstrate that the synchronized $SM/M/1$ queue can fill the gaps. The proposed queue theoretic method has no relation to the Cobb-Douglas production function. We also replaced 'Cobb-Douglas equations' with 'Cobb-Douglas production function' to reduce any confusions.

  The revised manuscript replaces 'balking' with 'probability of empty vehicles successfully picking up passengers,' which makes the proposed structures easier to understand. As shown in Fig. 5 in the revised manuscript, only empty vehicles who can successfully pick up passengers (in one specific spatial unit) can enter the corresponding taxi queue and match with one passenger. All other vehicles join the corresponding road queue and are processed to travel through that spatial unit. In the case study, the probability is also examined under proposed aggregation scales and study periods. We summarize the probability of choosing ATS in Fig 17. In specifics, Fig. 17(a) shows slightly higher pickup probability in downtown Manhattan, which is in line with official statistics. Fig. 17 (b) presents the variance of such probability across minutes. It is apparent that most spatial units are with very small variance of less than 0.03. All statistics support the assumption of fixed pickup probability by Uber drivers during peak hours. Similarly, we



(a) Mean (b) Variance

Figure 17: Probability of Uber driver partners successfully picking up in every minute and spatial unit during peak hours

summarize statistics for the case of off peak hours and TTS. We have similar findings, but do not show detail results here.

- Regarding road subsystem performance, there appears to be an important parameter c, defined in the nomenclature and the number of servers in the road subsystem and on page 8 as the critical taxi accumulations. Both need further explanation and dont look like the same thing. In Fig. 7, why there are multiple μ_k? Apart from being too small, this figure is unclear.

  Authors' response:

  The term 'critical taxi accumulation' is introduced to derive the number of servers in road queues. Here, we only focus on the interactions between taxi system and road congestion. We would like to measure at what accumulation level of taxi vehicles, the outgoing flow of one specific spatial unit will peak. For example, in the macroscopic fundamental diagram, once the number of taxi vehicles exceeds the 'critical taxi accumulation,' the congestion delays emerge for taxi vehicles within that spatial unit. Considering taxi's as probe vehicles, it can be seen that the 'critical taxi accumulation' is the same as c in the proposed road queue. Once the vehicle arrivals to the road queues exceeds the corresponding number of servers c, waiting queues emerge and a portions of the vehicles should wait for available servers. Additional clarifications have been added to the revised section of 'Inclusions of Road Network Performance.'

  Fig. 7 has been improved and now is Fig. 5 in revised manuscript. Also, μ_k has been replaced with μ_i^r, where i is the spatial unit and r refers to the fact that the service rate corresponds to the road. That is, the μ_i^k indicate the service rate of each server in road queues. This study models a road queue with a multi-server queue $M/M/c$ and assumes same service rate μ_i^r for each server. This assumption is justified because the urban road network can process vehicles in parallel and there are no differences in mobility efficiency among various service types and vehicle status.

- In Theorem 1, a variable x appears for the first time. This should appear in the nomenclature. Theorem 2 is without a proof.

  Authors' response:

The variable x indicates the system sate or the number of vehicles in the system. The subscript and superscript denote the spatial unit, vehicle status, and service types. We have added this into Nomenclature.

Theorem 2 is an extension of Theorem 1, expanding the stationary state from subsystem to the whole queueing network, and a classic one in the literature on Jackson networks. The proof can be straightforwardly completed, since the routing process over network is based on a fixed probability matrix or a Bernoulli splitting process. Thus, we do not provide detail proof for Theorem 2. However, we add one note that the proof can be easily extended from presented proof for Theorem 1.

- I am not sure what the purpose of the Case Study is. Is it a model verification, in which case what real world properties has the model reproduced? How do we know that the model captures the dynamics of both the road and taxi system? Or is this a study of New York, in which case what have we learnt about New York and are these findings generalisable?

  Authors' response:

  The case study is designed for a model verification in two separate hours - one peak hour from 6 to 7pm and an off peak hour from 10 to 11 am. The high-resolution mobility dataset from both TTS (yellow taxicab) and ATS (Uber) in New York City provides empirical observations on passenger arrivals, vehicle arrivals, modal split probabilities, probabilities of successfully picking up passengers, vehicle routing probability matrix, vehicle system-exiting probabilities, vehicle passenger-searching time, and vehicle trip duration traveling through one spatial unit. These observations are useful, not only as model inputs, but also for model validation.

  There are two inferred performance metrics presented in the revised manuscript: the approximated $\lambda_i^{pv,*}$ denoting paired vehicle-passenger flow arrivals in the synchronization process, and the sojourn time (i.e. total time from arrival to departure in one specific queue system). The $\lambda_i^{pv,*}$ can be estimated with the solution of proposed linear programming (equation 24 in the revised manuscript) under our proposed modeling structure. In reality, it generally corresponds to passenger pickups, observable with taxi trip records. The sojourn time inference is based on the queueing network solution. We split the full dataset into two parts (70% vs. 30%). The huge simulations based on the 70% of dataset lead to a reliable estimation on service rate of $M/M/1$. Then we use service rate estimation and 30% of dataset as inputs, thus obtain expected sojourn time in one specific spatial units. The empirical results are presented in revised section 4.3. The model outputs are comparative to those corresponding values observed from 30% of dataset, for most spatial units. In addition, we also admit that the proposed modeling structure performs relatively not very well for TTS in remote spatial units, due to limited empirical observations and imbalanced distribution of TTS activities. Need to revise this entire paragraph. I'm not sure what is being said here.

  The proposed modeling structure and the statistical methodology do not involve any conditions that are *especial* to New York City. Thus, the methodology and model can definitely be generalized to other instances. However, we should be cautious while transferring empirical findings on the spatio-temporal aggregation scale. Taxi activities may vary greatly from city to city due to different socioeconomic conditions, land use, demographics, and so on. Whether these empirical settings are comparative across instances is a very interesting problem but beyond the scope of this study.

Response to Reviewer 3

- The paper formulates stationary (or steady-state) queueing models to describe a road system that is shared by both traditional- and app-based taxis. Of course, these models assume that traffic conditions are in steady-state, which would be problematic in most any highway system. Yet,

Im not sure that much (if anything) is said about this heroic assumption. Section 3 discusses approximations, presumably to accommodate what the authors refer to as a synchronization process. My concern pertains to long relaxation times that occur in traffic, particularly during rush periods. Is this what the authors mean by a synchronization process?

Authors' response:

First, we would like to clarify the synchronization process. In our study, each defined homogeneous spatial unit (or subsystem) is modeled with two queues, one of which is the synchronization process addressing passenger and empty vehicle matching, and another that is about the traffic dynamics on the road network. Hence, the synchronization process is **not** related to the traffic movement over the road network and only describes the process of generation of empty vehicles and passengers to passenger pickups.

In specifics, the proposed synchronization process has two input flows that are empty vehicle arrivals (can be vehicles either newly online or transferring from other spatial units) and passenger arrivals. The homogeneous spatial unit as one taxi queue matches them together with zero-matching time once they generate and processes matched pairs with a service rate depending on region configurations. The approximations in section 3 further discuss the asymptotic properties of such synchronization process with two input flows and matching behaviors. This indicates that once the observation time t goes to infinity, the synchronization process can be approximated by a standard $M/M/1$ taking the minimum of passenger and vehicle arrival rate as new one and keeping original service rate. This significantly reduces difficulties in stationary state analyses over networks of synchronization processes (generally with unstable system states).

Second, we can show that the queueing network can reach a stationary state (or 'relaxes') and may be true in reality. On one hand, our queueing network is built upon one-hour peak (or off peak) hour. In such short duration, traffic conditions can assume to be in steady-state. On the other hand, although we lack empirical evidence of how long the relaxation time is, this has been a subject of long-standing discussion in queueing theory. This literature provides an upper bound, $t < ((1+\rho)/(1-\rho))^2$, to measure this relaxation time[5]. Based on this recommendation, our two case study's (peak hour and off peak hour) can relax within an observation period when the utilization rate is less than 0.95. Most spatial units meet such requirements. Only one spatial unit in midtown Manhattan has higher utilization rate of more than 0.95 during peak hour, regardless of whether the service is ATS or TTS. This spatial unit (with utilization of 0.97) demands more observation times to relax. During off peak hours, there are no spatial units with utilization more than 0.90 for either service types.

- Because of my concern, I was especially interested in the case study presented in sec. 4. But I find the presentation unsatisfying. The discussion in sec. 4.1 seems muddled to me. The difference in market size is said to be due to the different years for the [ordinary taxi] and [app-based taxi] data. What difference in market size? I see no data pertaining to this. Another example: There is said to be 366K and 539K empty trips. What constitutes an empty trip? It is further said fleet size and total service hours of taxi cabs do not change significantly Presumably this means that the values do not change significantly over time. But what is the time scale? Are we talking about changes across a day, across distinct days or across years? Why does this even matter? Traffic conditions change with time, whether or not taxi fleet size changes. On a (possibly) related note, the authors claim to have identified the critical taxi accumulations resulting in road congestion. But a major part of road congestion has to do with the accumulations of privately-owned cars. Is their contribution to congestion somehow not important?

Authors' response:

Section 4.1 primarily presents the data source and case study development. We obtained the yellow taxicab (one typical TTS) trip records from New York City in 2017. From this dataset it is easy to estimate passenger arrivals. However, there are no vehicle id's attached to each

trip record, and this increases the difficulty in finding sequential trips by a single vehicle. As a consequence, we cannot get any information on TTS vehicle routing and system exiting behaviors. Alternatively, we infer such missing information based on yellow taxicab trip records in anther year of 2013. Although there are slight differences in TTS pickup distribution between 2013 and 2017, we think the 2013 estimation on vehicle routing and system exiting behaviors can be a reliable approximation for those in 2017. This is because of almost no changes in taxicab fleet size and shift hours from 2013 to 2017.

Last, we only focus on the interactions between taxi system and road congestion. We would like to measure at what accumulation level of taxi vehicles (rather than bus and autos), the outgoing flow of one specific spatial unit peaks. For example, in the macroscopic fundamental diagram, once number of taxi vehicles exceed the 'critical taxi accumulation', the congestion delays emerge for taxi vehicles within that spatial unit. One might consider taxi's as 'probe' vehicles, and we refer to the above accumulation notion for determining the number of servers in $M/M/c$ and capture congestion delays. Note that, taxis as probe vehicles are widely used in macroscopic fundamental diagram derivation methods. In conclusion, the observed congestion experienced by taxis in the network is a consequence of the accumulations of privately-owned cars, buses and other vehicles. Indeed, their contribution is significant, and not something that is being ignored.

The section is also revised carefully with empirical support and above arguments.

- Section 4.2 is titled Queue Settings, and seems to entail parameter estimation. Estimated arrival rates are said to indicate the unbalanced distribution of taxi services over space How does the reader see this in the data? On a separate note, Fig. 8 indicates that the assumed distributions of inputs tend not to hold. Is this not a concern?

  Authors' response:

  Section 4.2 is revised in the following two ways:

  First, the passenger and vehicle arrivals are extensively examined with hypothesis testing methods, for instance Kolmogorov-Smirnov (KS) test and three additional χ^2 distribution based tests. These methods are developed to check whether arrival counts can be described with one (homogeneous) Poisson distribution. In more than 70% of spatial units, the arrivals are thought to be consistent with Poisson distribution. We present hypothesis testing results in this section. In addition, we summarize statistics of pickup and system exiting probability, assumed in the queueing network. We also obtain empirical support for assuming fixed values in each spatial unit.

  Second, we also include multiple additional plots for spatial distribution of input values, including arrival rates, pickup probability, and system exiting probability. The spatial distribution plots clearly indicate imbalanced taxi activities over space, for instance arrival rates shown in Fig. 9 (d) to (f) in the revised manuscript.

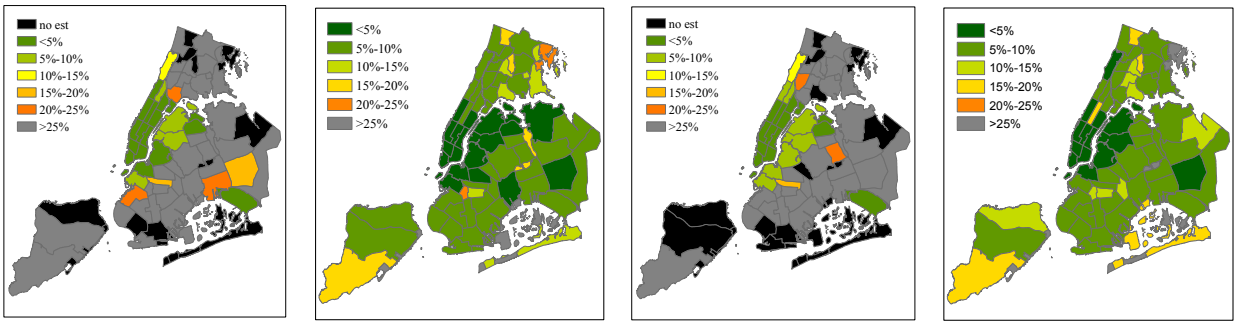
- The evaluation of the models in sec. 4.3 is quite troubling to me. Figure 11 indicates that differences between empirical and predicted outputs often exceed a factor of 2. Explanations offered by the authors pertain to an unbalanced distribution of taxi activities, but the discussion seems muddled to me. I wonder if the assumption of steady-state conditions is playing a role here, especially since the greatest discrepancies occur during peak periods.

  Authors' response:

  Due to changes in derivation for service rate (section 2.2), we also change the model estimation method and outputs accordingly. There are two properties reproduced in the revised manuscript, including approximated $\lambda_i^{pv,*}$ denoting paired vehicle-passenger flow arrivals of synchronization process, and sojourn time (i.e. total time from arrival to departure in one specific queue system). The $\lambda_i^{pv,*}$ generally corresponds to passenger pickups in reality, which is also solution

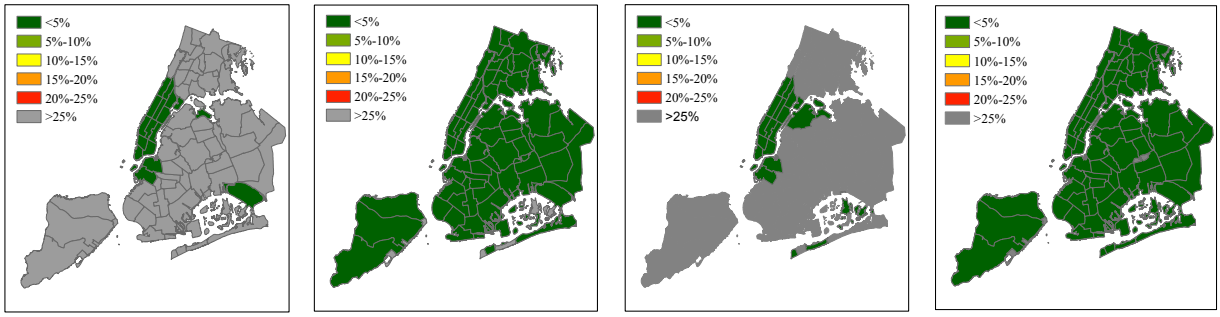
of proposed queueing network. The second reproduction of sojourn time is based on queueing network solutions, rather than direct calibration on unknown queue service rates. We split the full dataset into two parts (70% vs. 30%). The huge simulations based on the 70% of dataset lead to a reliable estimation on service rate of $M/M/1$. Then we use service rate estimation and 30% of dataset as inputs, thus obtain expected sojourn time in one specific spatial units. The empirical results are presented in revised section 4.3 and two important findings are also shown here.

Fig.18 and 19 show almost same patterns between peak and off peak hours, but reveals many differences in model accuracy between ATS and TTS. The ATS system presents much lower absolute percentage errors (i.e. <5%) in almost every spatial unit. In contrast, the proposed modeling structure has reliable outputs for "hot" areas of TTS system, where attract more than 90% TTS activities in reality. Such significant differences may arise from spatial distribution of both services. Since the modeling structures involves vehicle movement over road network and routing probabilities, which directs majority of vehicles to "hot" areas and leads to unreliable estimations for remaining areas. In addition, modeling outer Manhattan areas (gray areas in figures) are limited by not only its highly imbalanced distribution but also very few empirical observations. We also added above explanations in the revised manuscript.



(a) TTS - peak hours (b) ATS - peak hours (c) TTS - off peak hours (d) ATS - off peak hours

Figure 18: The mean absolute percentage errors between expected sojourn time and observed one at taxi queues



(a) TTS - peak hours (b) ATS - peak hours (c) TTS - off peak hours (d) ATS - off peak hours

Figure 19: The absolute percentage errors between estimated $\lambda_i^{pv,*}$ and observed passenger pickup flows

- On a different note, the authors claim that traditional- and app-based taxis have similar door-to-door mobility services, and the present models seem to reflect this notion. In reality, app-based

taxis provide an array of services, and as a result a taxi can at any one time have multiple passengers on board with distinct trip origins and destinations. Perhaps these distinct services can be ignored in the name of approximation, yet no mention of this matter is offered in the paper.

Authors' response:

The app-based taxis provide multiple differentiated mobility products, varying from economy (e.g. UberX), shared (e.g. UberPool), premium (e.g. UberBlack), special (e.g. UberWAV and UberFamily), to delivery. Our study is limited to the economy service and does not include any other app-based taxi products. The described app-based taxis in this comment is shared mobility, which only has a small portion of trip requests and is just a pilot program in several cities. It is out of the scope of the current paper. In addition, we add one short note in first paragraph to clarify taxis of interests.

- Overall, I think that the quality of the presentation is poor. Terms are often introduced without definition. In some cases, definitions come later in sec. 2.1, but I find the discussion in this section to be muddled as well. For example, it is said that [d]ata science approaches tend to be process agonistic Do the authors mean agnostic? What does it mean to predict the growth of system latency? Other statements elsewhere in the paper make no sense to me. Why, for example, would optimal control be critical to understanding the market dynamics [emphasis added]? There are numerous grammatical errors in the paper, and these do not help the situation.

  Authors' response:

  The listed terms are mainly in section 1 'Introduction'. We rewrote the second-to-last paragraph in section 1 and improves the corresponding expressions. For those uncommon terms, such as 'agnostic' and 'predict the growth of system latency', we removed them to make corresponding sentences easier to understand. In addition, we proofread the revised manuscript and corrected grammar errors, as well as typos. Hopefully, the reviewer finds the paper to be an easier read now.

References

[1] Xinwu Qian and Satish V. Ukkusuri. Taxi market equilibrium with third-party hailing service. *Transportation Research Part B: Methodological*, 100:43 – 63, 2017.

[2] Hai Yang, Cowina W.Y. Leung, S.C. Wong, and Michael G.H. Bell. Equilibria of bilateral taxi-customer searching and meeting on networks. *Transportation Research Part B: Methodological*, 44(8):1067 – 1083, 2010.

[3] CONSTANCE L. WOOD and MICHELE M. ALTAVELA. Large-sample results for kolmogorov-smirnov statistics for discrete distributions. *Biometrika*, 65(1):235–239, 1978.

[4] Lawrence D. Brown and Linda H. Zhao. A test for the poisson distribution. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 64(3):611–625, 2002.

[5] J.P.C. Blanc. *On the relaxation time of open queueing networks*, pages 235–259. Number 7 in CWI monographs. North-Holland Publishing Company, 1988. Pagination: 25.