A comparison of taxi usage patterns in New York and San Francisco using LDA

Lisa Rayle | December 2014 | Final project for CE263

MOTIVATION

The recent rise of Uber, Lyft and similar services has turned attention to on-demand urban transportation, a role which has long been filled mainly by taxis. Few doubt Uber et al. are well on their way to replacing some, if not most, of the taxi and livery industry. More in question is the extent to which these services, by expanding the availability and increasing reliability of on-demand rides, may be expanding the traditional market. In cities like New York, where the traditional taxi ecosystem is well developed, Uber may represent only a marginal change. However, in cities with sparser taxi service, like San Francisco, Uber may serve a wider market than do taxis.[1] Understanding these changes, if there are any, requires a baseline understanding of the previously existing taxi market. Few studies have researched taxi usage in these cities and, to my knowledge, no comparative studies exist.

QUESTION

In this project, I compare temporal and spatial characteristics of taxi usage in New York City and the San Francisco Bay Area using GPS data from taxi cabs. Based on the time and location of pick-ups and drop-offs, in each city I employ Latent Dirichlet Allocation (LDA) to identify characteristic "types" of trips. I expect to find that New Yorkers tend to use taxis in a broader range of cases, especially for commuting, whereas taxi usage in San Francisco is more specialized, with a greater focus on late night and airport trips. The use of LDA would ideally provide a way to infer taxi trip purposes from GPS data alone, without need for surveys or reliance on potentially biased data sources.

DATA SOURCE AND PREPARATION

This analysis uses timestamped taxi GPS records provided by the SFMTA[2] and the NY Taxi and Limousine Commission.[3] The San Francisco dataset consists of complete trip records (more than 700,000 total) from one of the city's larger taxi companies, DeSoto Cab Co., for October 2012 and mid-July through October 2013. The New York dataset includes all trips in the city. I chose to use only October 2013, which totals more than 15 million trips[4]. Each dataset includes coordinates and timestamps for each trip's origin and destination. For New York, I randomly sampled 10% of the data. After dropping records with missing data or lat/long coordinates lying far outside the metropolitan area, I obtained a set of N=~1.4 million for New York and N=702,032 for San Francisco. The data cover the

[1] Noting, of course, that smartphone taxi hailing apps like Flywheel are blurring the distinction between taxis and Uber.
[2] Obtained through personal communication.
[3] See http://chriswhong.com/open-data/foil_nyc_taxi/
[4] I chose October because it has no major holidays and typically few weather extremes, and it can be compared with the SF dataset. The SF data includes summer months and so might reflect more tourist activity.

metropolitan area of each city. The data are from 2012 and 2013, a time when Uber was only beginning to impact the taxi market.

To define the origin and destination locations, for each city I created a grid of 1000m x 1000m cells. Each OD pair thus corresponds to an ordered pair of grid cells. I also considered using a 500m grid, census tracts, or traffic analysis zones to define origins and destinations. The 500m grid was too small for San Francisco--there were too few trips in each cell--and while it was better for denser New York, I wanted consistency between cities. For future New York analyses, a 500m grid might be more appropriate, especially to identify patterns within Manhattan. The current 1000m grid is good for analyzing patterns at metropolitan-level and between broad sections of Manhattan. Census tracts, while an appropriate size, proved to be poor definitions because tract boundaries typically lie on streets, exactly where trip observations occur. TAZs might seem a good candidate because they are defined specifically to represent trip generating locations, however I was unable to find consistent TAZ definitions for New York.

I created the grid shapefile in QGIS and then used PostGIS/Postgresql spatial queries to join each trip origin and destination with its corresponding grid ID.

THE LDA MODEL

LDA is a probabilistic generative model originally developed to analyze document collections, as described by Blei et al. (2003). The central idea is that each document in a given collection contains a mixture of latent topics, and these underlying topics give rise to a predictable vocabulary. Thus the pattern of word frequency in the document collection can be used to infer the topics it contains.

Following Blei et al. (2003), the model is formalized as follows. Given the parameters α and β, the joint distribution of a per-document topic mixture θ, a set of per-word topic assignments $\mathbf{z}$, and a set of N words $\mathbf{w}$ is:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta).$$

The topic mixture θ is drawn from the Dirichlet distribution ~Dir(α) and thus the parameter α governs the shape of per-document topic distributions. The parameter β represents the probability of choose a word given a topic, or $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, and is to be estimated. Figure 1 shows a graphical model of LDA.

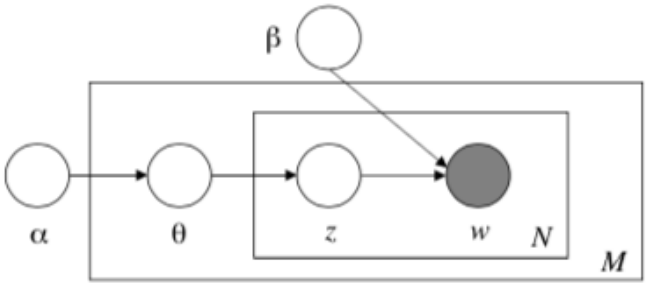


Figure 1: Graphical model of LDA. The observed word "w" is drawn from a topic z, which is drawn from a topic distribution θ. *(Source: Blei et al., 2003)*

LDA APPLIED TO TAXI TRIPS

This project applies LDA to the analysis of taxi trip patterns. Others have adapted LDA to spatial contexts such as bikeshare trips (Come et al., 2014) and urban activity patterns (Kling and Pozdnoukov, 2012). In this application, each "word" is an OD pair and each "document" is the set of trips occurring within a one-hour period. The "corpus" is all the trips over all time periods in the city. The inferred "topics" can be thought of as "types" of trips that might correspond to trip purposes or activities. Come et al. (2014) refer to topics as "OD-templates" from which trips are drawn. Here I will use "topic" to avoid confusion (though I agree that "template" is perhaps a better description).

I used the LDA implementation in the Gensim Python package. To process the data, I grouped trips into one-hour periods, obtaining 744 "documents" of roughly 1000 OD pairs each for New York and 3696 documents of about 100 pairs each for San Francisco. As is typically done in LDA, I removed pairs that occurred in only one document, leaving approximately 20,000 and 30,000 unique pairs for San Francisco and New York, respectively.

To tune model parameters, for each city I held out a portion of documents corresponding to representative days as a test set. I then trained the model on the remaining data. I selected the parameter α by minimizing the perplexity of the test data. Optimal values were $\alpha = 0.25$ and 0.5 for San Francisco and New York, respectively (Figure 2).

In text analysis applications, the number of topics k is often chosen by perplexity minimization as well. In this case, however, the goal is to identify interpretable topics. Perplexity should generally decrease with increasing k, so the idea is to choose the lowest k that still gives meaningful results. After trying a range of k=3 to 6, I decided the most interpretable results were with k=5. I ran the LDA algorithm several times to ensure the results were stable. For purposes of comparability, I used the same k value for both cities, although, as discussed later, different values may be more appropriate.

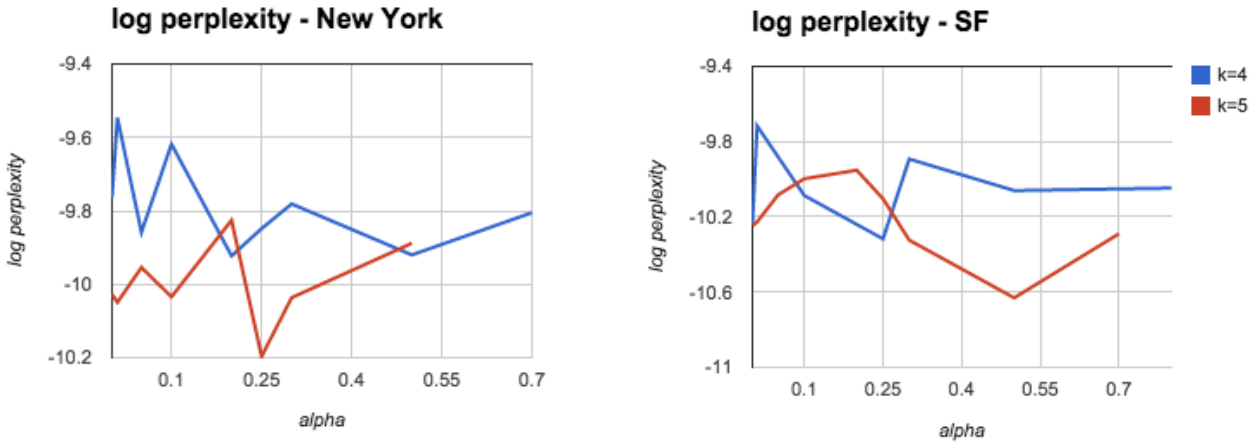


Figure 2: Choosing parameters through perplexity minimization

RESULTS

The LDA models appear to describe the data fairly well. To visualize the appearance of topics over time, I multiplied the posterior probabilities by the number of trips observed for each day. Figure 3 shows the expected number of trips generated from each topic for each hour over the first full week in October 2013. One can see the clear dominance of certain topics at particular times of day.

The maps in at the end of this document show the probabilities of trip origins and destinations for each topic.

From these figures, it is possible to qualitatively describe the topics. In LDA, the ordering of topics is arbitrary, thus this list does not imply any order of importance.

For San Francisco:

- **Topic 1 (light blue): "Recreation/other"**
  Applies very rarely and appears to represent ODs not common in other topics. It exhibits more dispersed pattern, with more leisure destinations (such as the Berkeley Marina) and covers destinations in the East Bay and the Peninsula that are unlikely under other topics.
- **Topic 2 (dark blue): "Tourist/airport"**
  Appears to be mainly airport trips, especially from Union Square. Most likely in the very early morning.

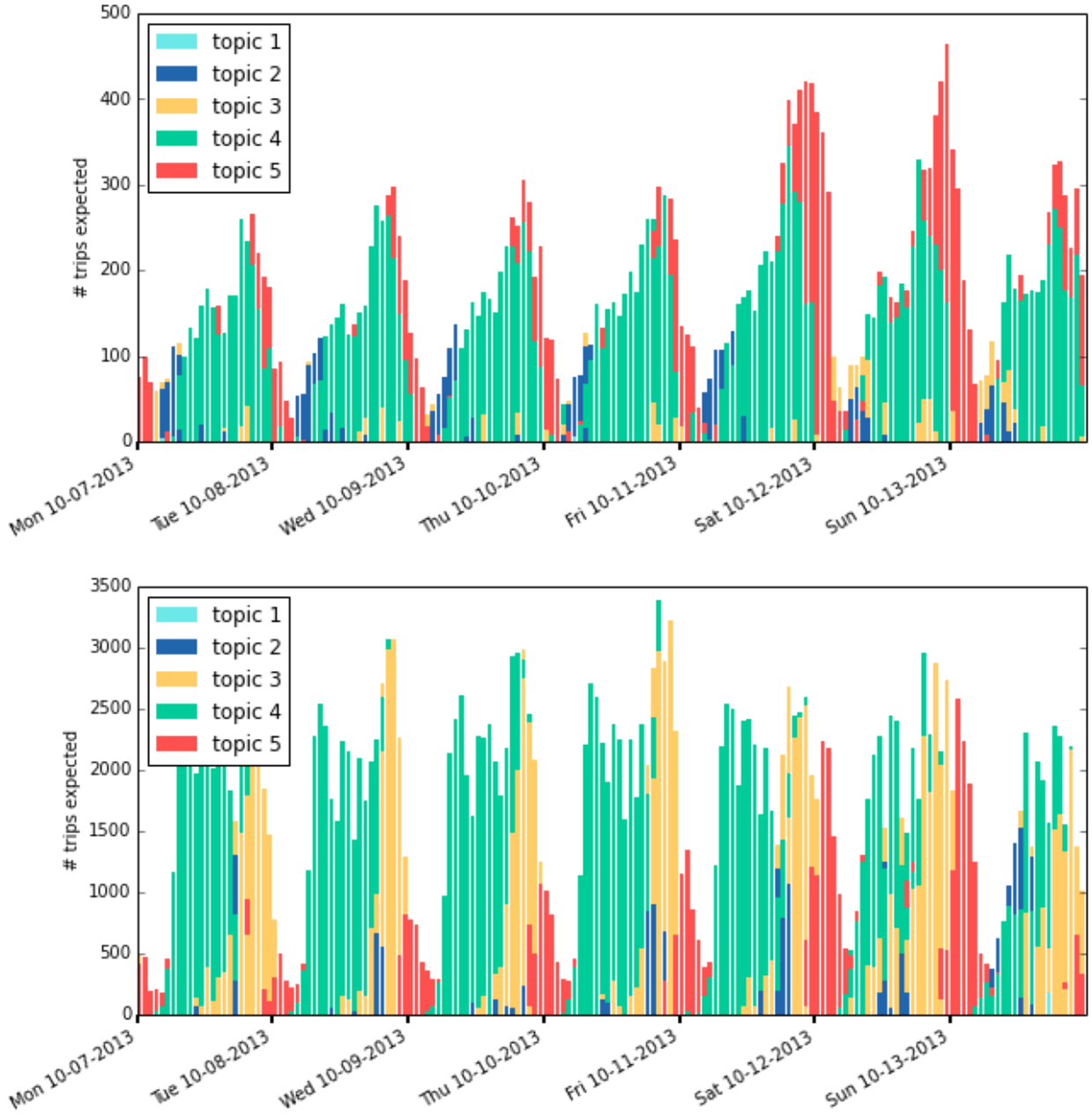


Figure 3: Number of expected trips from each topic by hour over a one week for San Francisco (top) and New York (bottom). The x-axis labels mark beginning of each day (i.e., just after midnight of the previous day.)

- **Topic 3 (yellow): "Late night/airport"**
  Most prevalent on Saturdays and Sundays around 2am-5am. Trips are likely to begin or end in late night spots like Soma, the Tenderloin, and Marina, along with the airport. It is difficult to distinguish from the late night and airport trips.

- **Topic 4 (green): "Work"**
  Describes trips during most of the day, with origins concentrated downtown and destinations downtown and extending down the peninsula along highway 101. This clearly represents work trips.
- **Topic 5 (red): "Night/social"**
  Applies to evening hours and especially Friday and Saturday nights. Origins are highly concentrated in a few locations--Tenderloin, SOMA, the Mission and the Marina. Destinations are likely to be those same locations but also other residential neighborhoods, indicating this topic largely represents trips between social destinations and trips home.

For New York:

- **Topic 1 (light blue): "Other"**
  Like Topic 1 for San Francisco, this rarely applies, and appears to be relatively dispersed across the city. It is probably a catch-all topic representing OD pairs that are unlikely under other topics.
- **Topic 2 (dark blue): "Airport/tourist"**
  Appears at midday and on Sundays. Origins and destinations are likely across Manhattan and at the LaGuardia and JFK airports. The highest concentrations are in midtown, where visitors to the city are likely to stay.
- **Topic 3 (yellow): "Evening commute"**
  The large spike in trips during the evening hours is drawn from this topic, which clearly represents the evening commute. Trip origins are highly concentrated in lower and mid Manhattan; destinations are more dispersed.
- **Topic 4 (green): "Morning commute"**
  Prominent during morning and midday. Origins are concentrated in financial district and around central park. Destinations are concentrated in midtown, as well as LaGuardia.
- **Topic 5 (red): "Night/social"**
  This clearly represents late night, social trips. It applies to late night hours, especially Friday and Saturday nights. Trips are very likely to originate in a handful of locations in the Lower East Side and Williamsburg, while they are likely end in the same locations as well as elsewhere in Brooklyn and Manhattan. Airports do not appear on map.

DISCUSSION

The results highlight some differences between taxi usage in the two cities. The first thing immediately noticeable from Figure 3 is that New York has many more morning taxi trips than San Francisco; the latter's trips are much more concentrated the evenings. The LDA analysis suggests there are distinct morning and evening commutes in New York, whereas in San Francisco we find no separate topic for the morning commute. This is probably not because five topics is too few, since Topic 3 already seems to be an "extra" topic. Even on weekends San Francisco trips are concentrated at night, compared to New York's more even distribution.

Another temporal difference between the two cities is that, as expected, San Francisco's "Night/social" topic begins earlier than New York's (Figure 3). In San Francisco, this topic applies to trips starting around 7pm, at the same time as the evening commute, compared with 10 or 11pm in New York.

For purposes of comparison, I used k=5 topics for both cities, but in San Francisco, there may be only four distinct topics. Topic 3 seems to capture both late night- and airport-type trips, based on its temporal and spatial probabilities. It may be distinct from Topics 2 and 5 in terms of the direction of OD pairs--this is something to consider in future analysis--but most likely it is simply an "extra" topic. In comparison, it was possible to clearly describe each of the five New York topics. This is evidence that taxis serve a richer range of uses in New York than in San Francisco.

CONCLUSIONS

In this project I used an LDA approach to identify the latent "topics" underlying observed taxi trips. The analysis supports the notion that the New York taxi usage arises from a wider range of trip purposes or activities than in San Francisco. In future work, it may be interesting to explore patterns at a finer spatial resolution, especially in New York, where sufficient data are available. Another extension could be to validate the trip purposes or activities using location information, for example, from social media data. The larger question motivating this project is the extent to which Uber et al. are expanding the market for on-demand travel. This same methodology could theoretically be applied to Uber trip data and would allow a longitudinal analysis and a comparison with taxi trips.

REFERENCES

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.

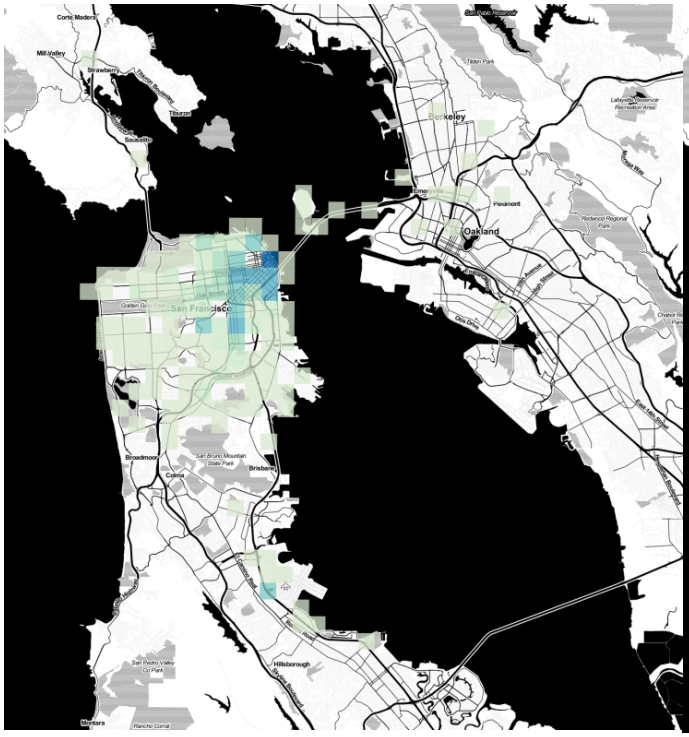
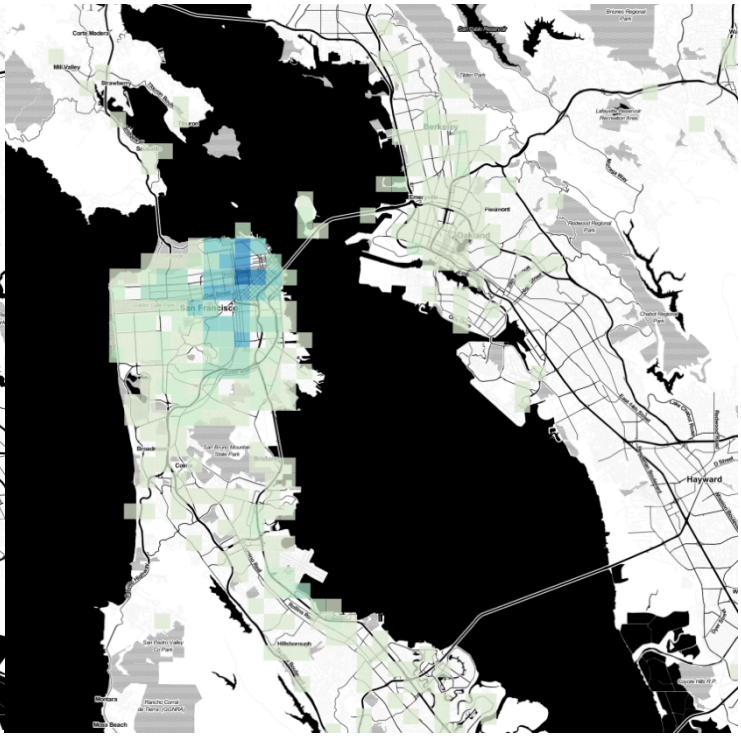
Come, E., Randriamanamihaga, N.A., Oukhellou, L., Aknin, P., 2014. Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation: Application to Vélib' Bike Sharing System of Paris, in: TRB 93rd Annual Meeting. Transportation Research Board, p. 19

Kling, F., Pozdnoukhov, A., 2012. When a City Tells a Story: Urban Topic Analysis, in: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12. ACM, New York, NY, USA, pp. 482–485.

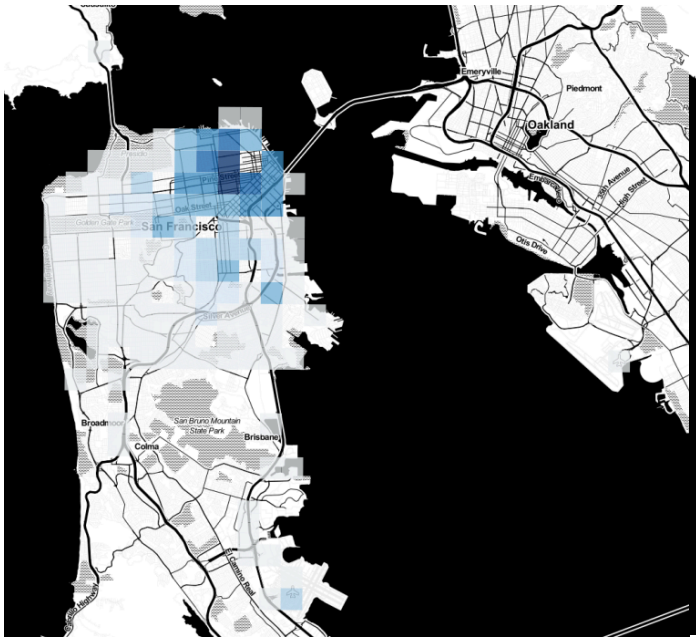
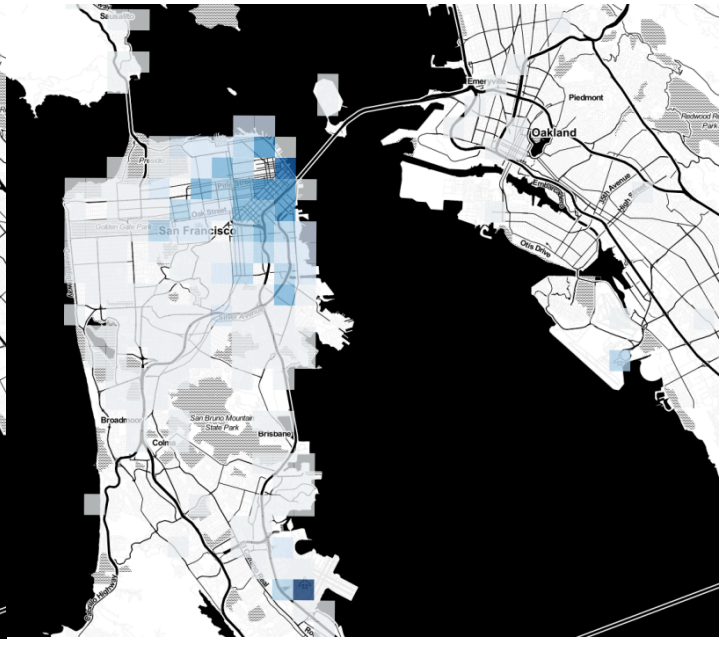
Probabilities of trip origins and destinations by topic

Shading indicates higher probability of a trip drawn from the corresponding topic originating or ending in that location.

San Francisco

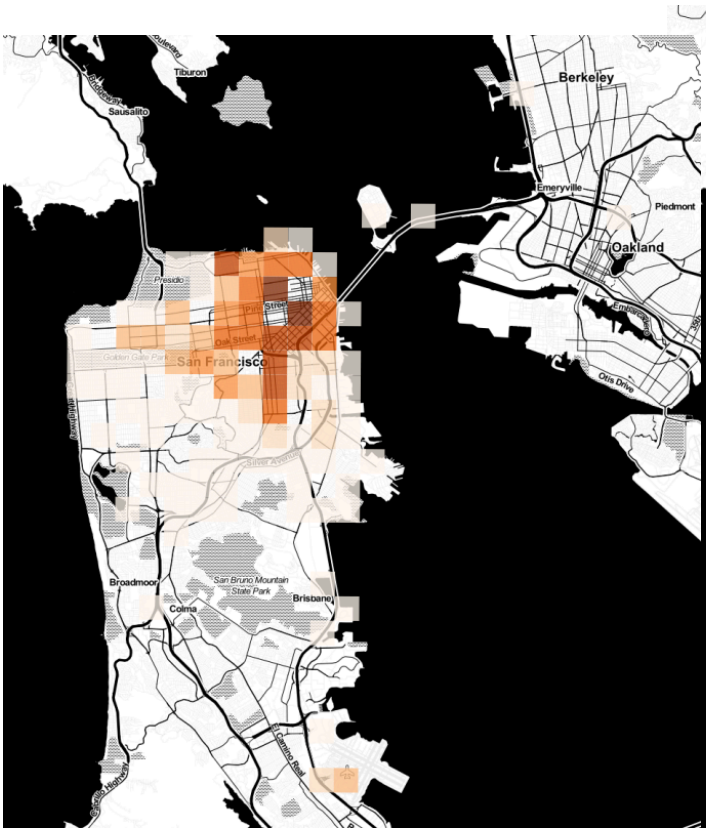



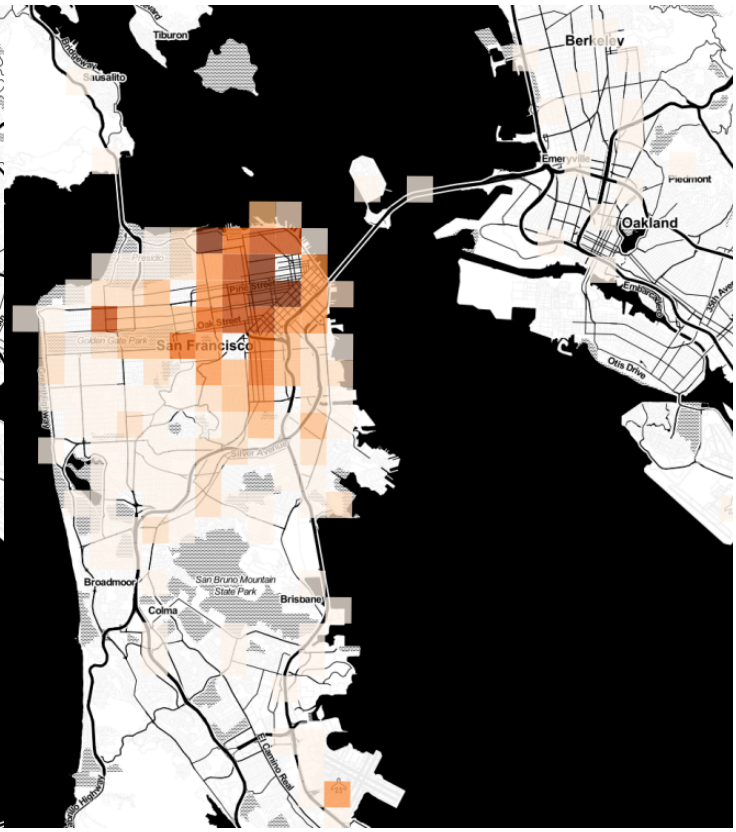
Topic 1: Recreation/other origins destinations

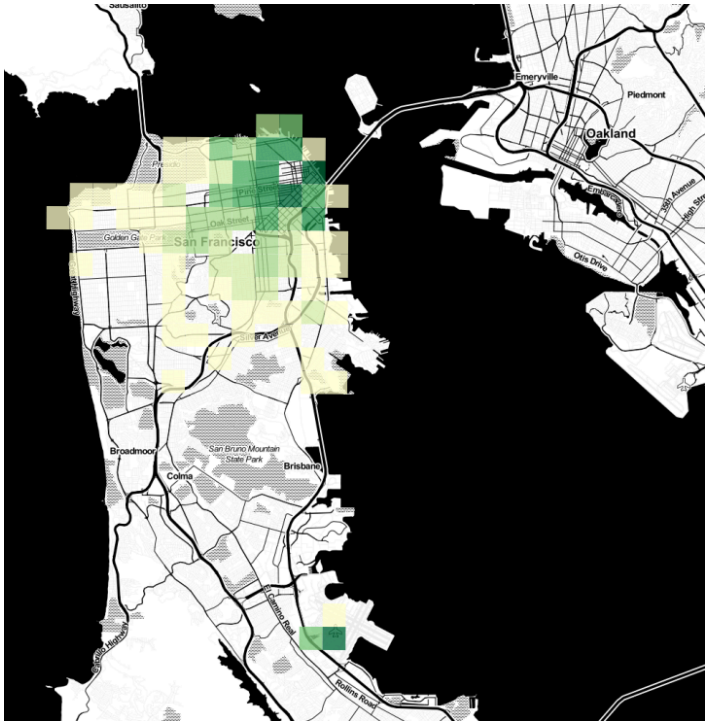
Topic 2: Tourist/airport origins destinations

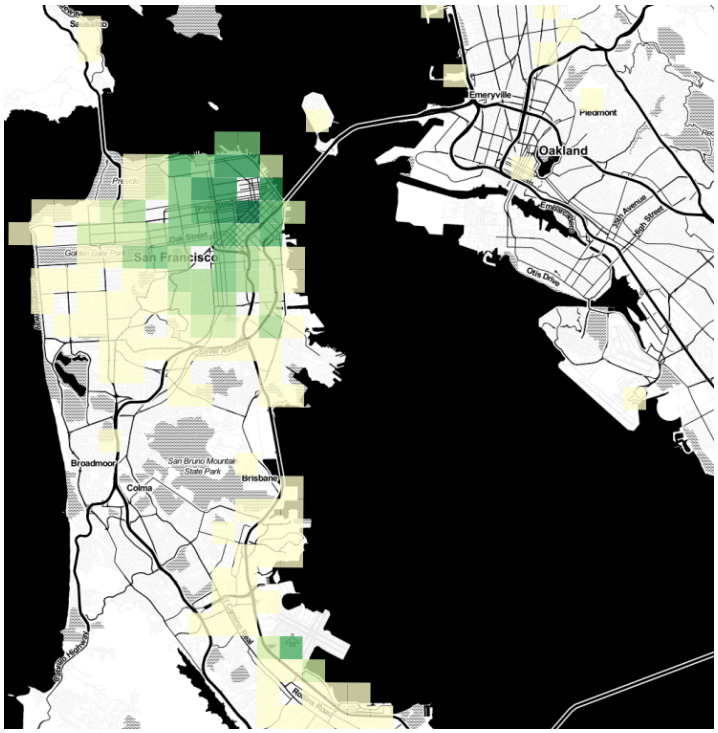
(map tiles by Stamen)



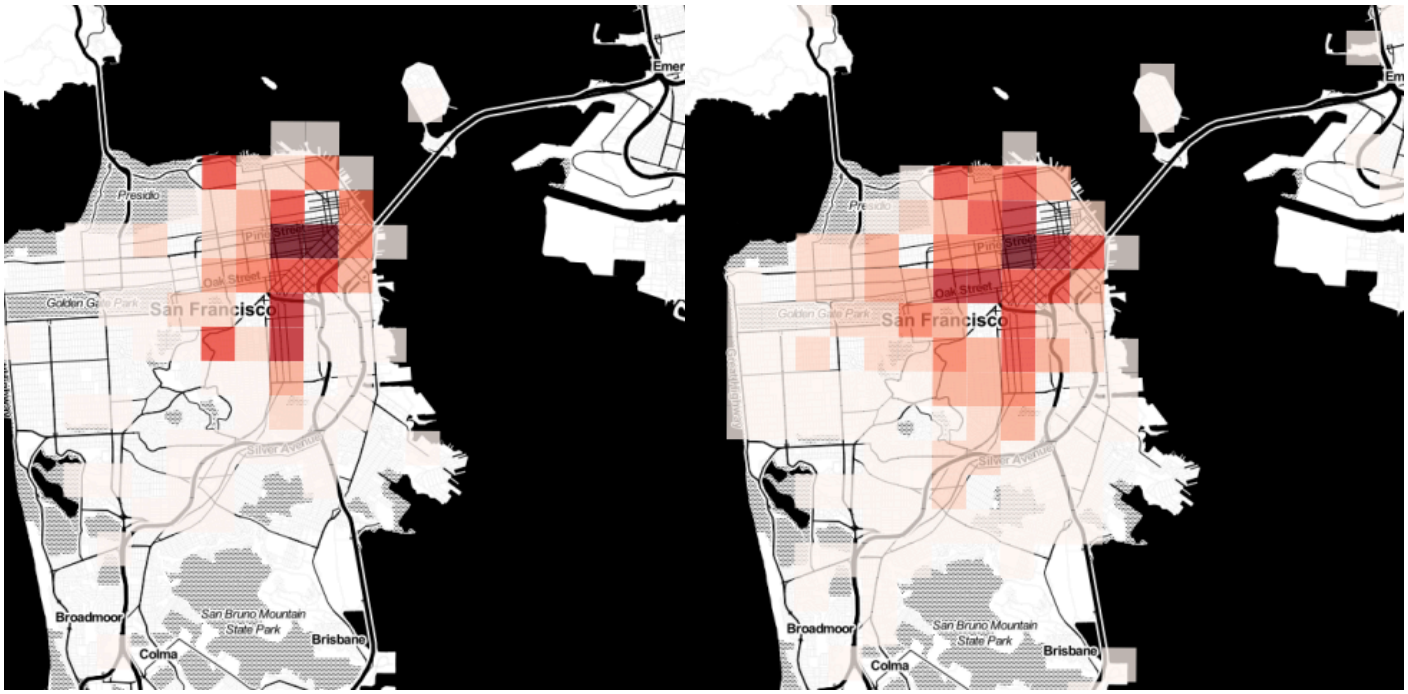


Topic 3: Late night/airport origins destinations



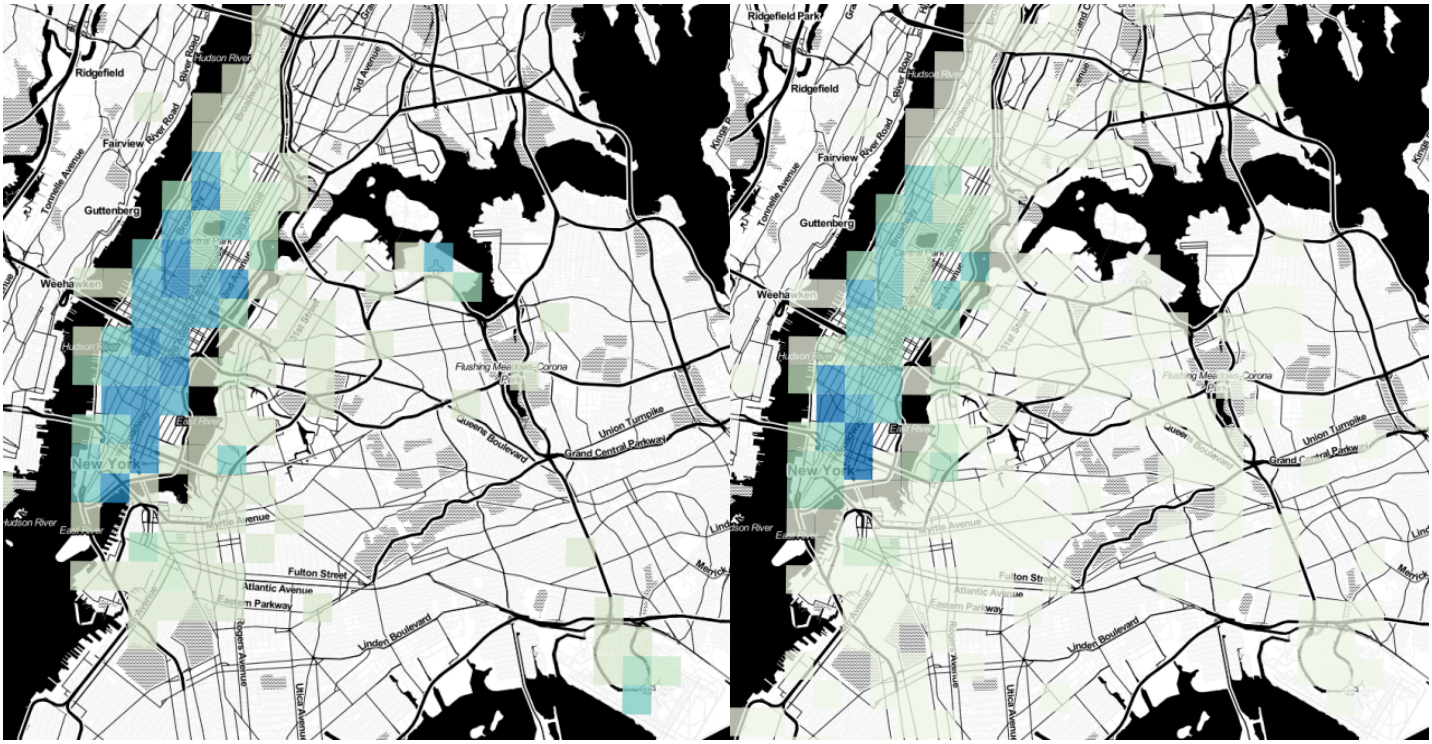


Topic 4: Work origins destinations

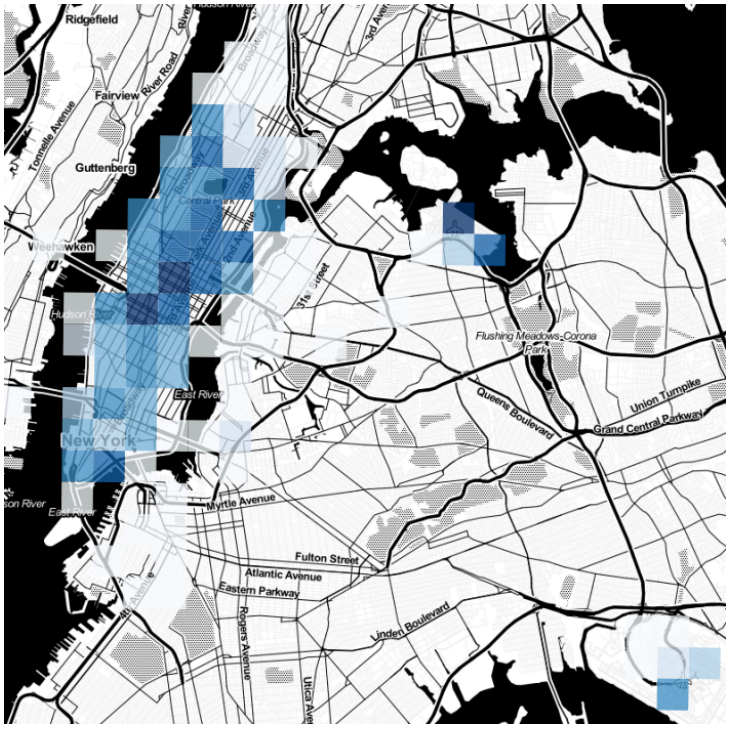


Topic 5: Night/social origins destinations

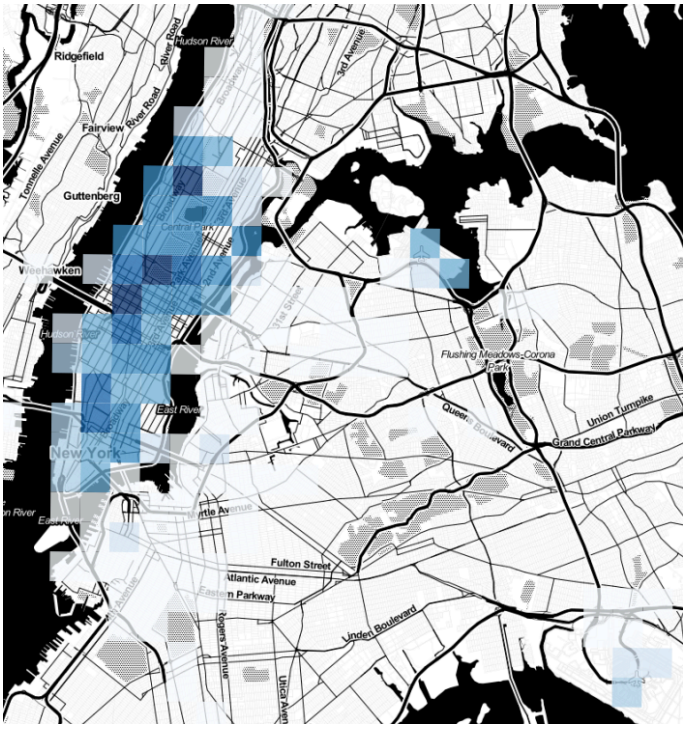
New York



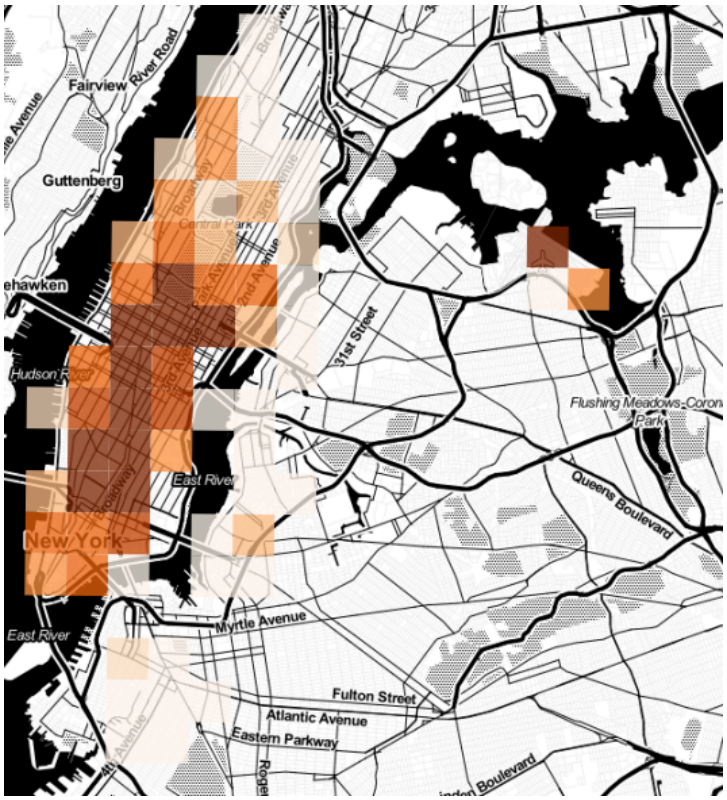
Topic 1: Other origins destinations



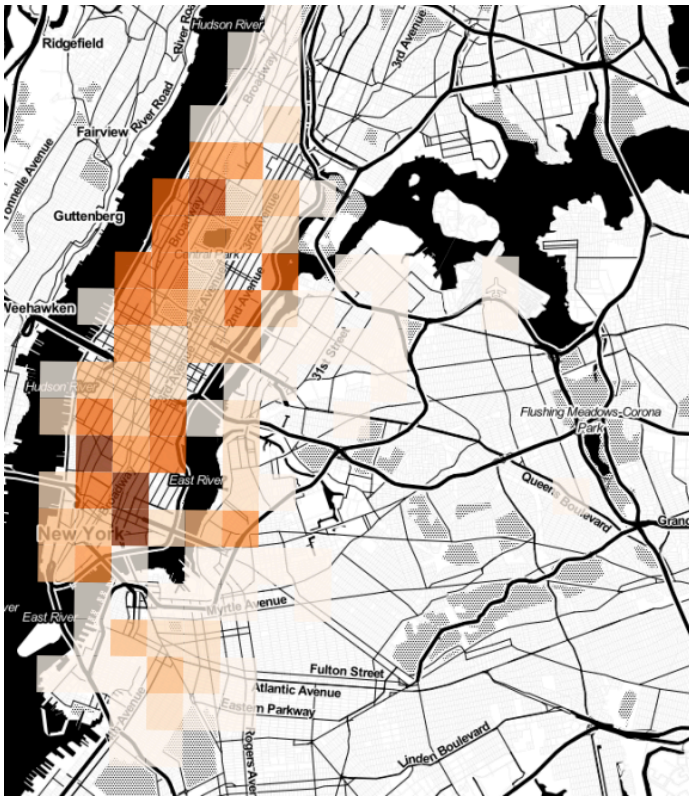
Topic 2: Airport/tourist origins



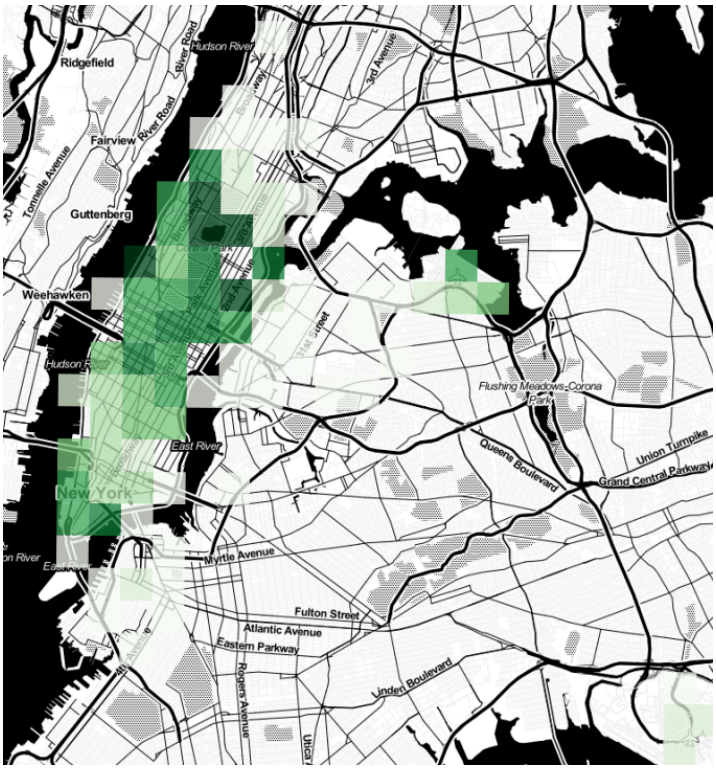
destinations

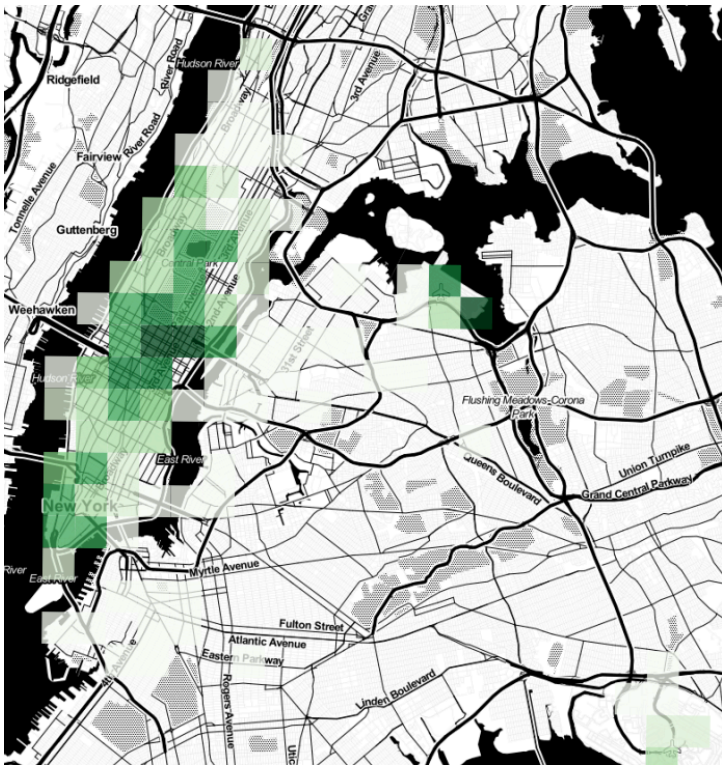


Topic 3: Evening commute origins

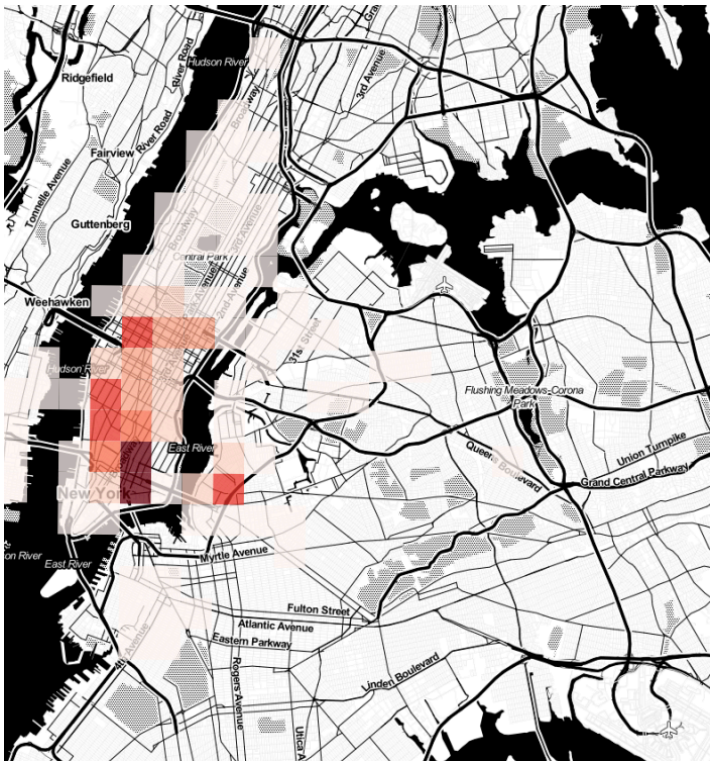


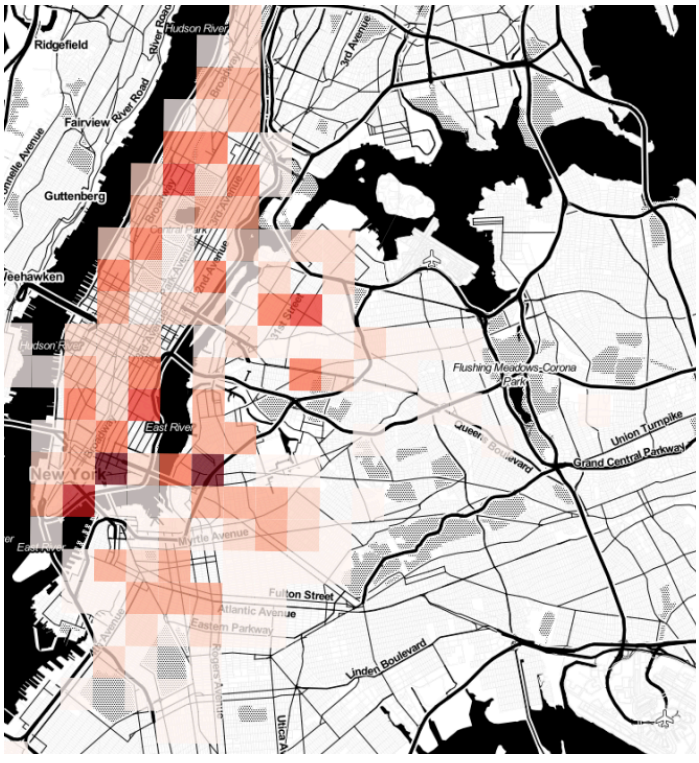
destinations





Topic 4: Morning commute origins destinations





Topic 5: Late night/social origins destinations