Dynamically Weighted Spatiotemporal Interpolation for Modeling Distribution of Twitter Population

Dandong Yin, Shaowen Wang

{dyin4,shaowen}@illinois.edu

CyberGIS Center for Advanced Digital and Spatial Studies
CyberInfrastructure and Geospatial Information Laboratory
Department of Geography and Geographic Information Science
University of Illinois at Urbana-Champaign

Location-based Social Media

- LBSM data are increasingly used to model population dynamics

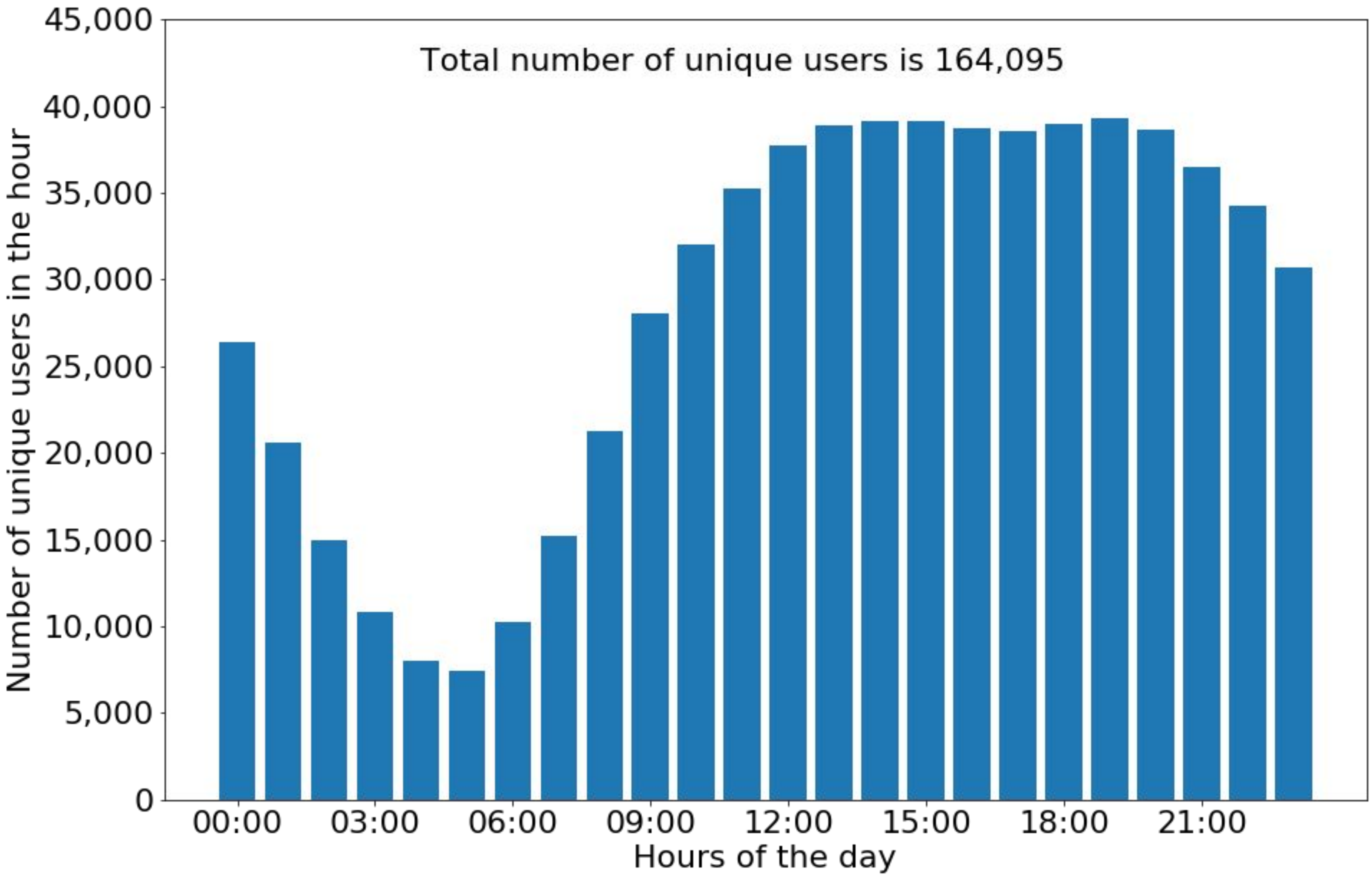
- Pros
  - Large volume
  - high resolution
  - real-time updates
  - easy accessibility
  - ….

- Cons
  - Sampling bias
  - Uncertainties:
    - Position
    - Continuity
    - Route

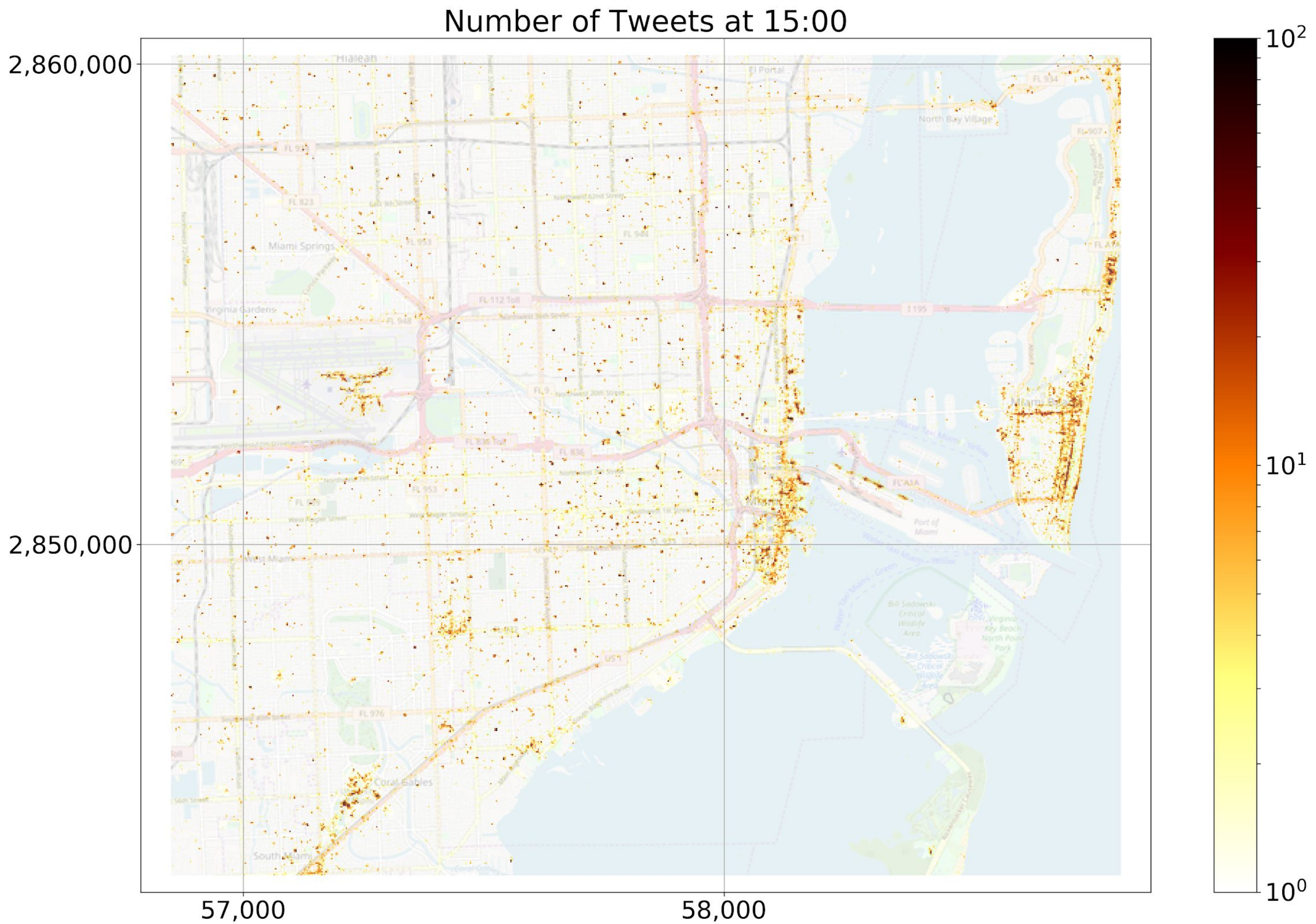
Sparsity and Uncertainty

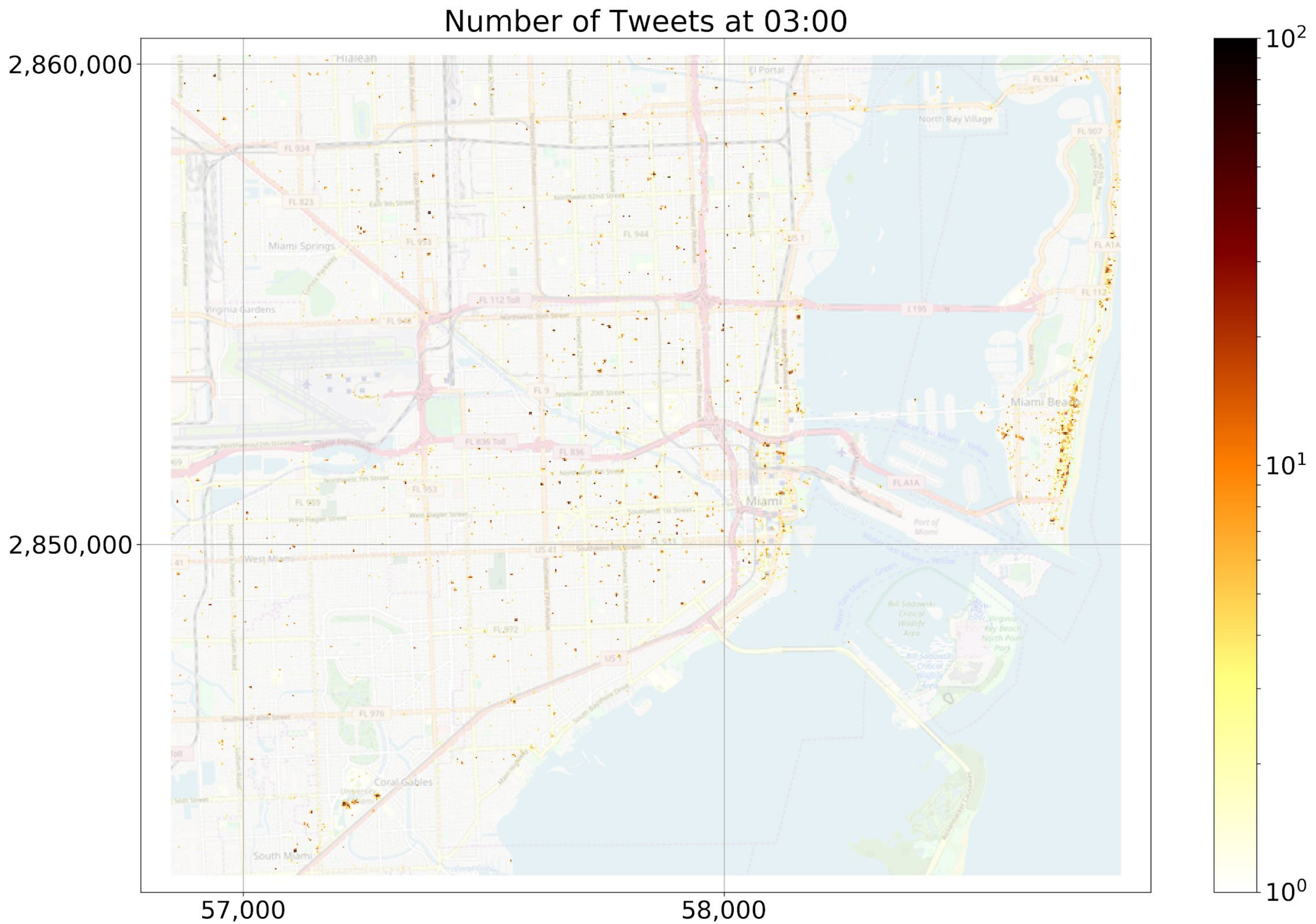
- Average tweeting frequency is relatively low, as compared with typical GPS-tracking data
  - The average spatiotemporal density of raw data records is quite sparse, and not evenly distributed
  - Inferring user activities between LBSM records is important for population modeling
  - A uncertainty-aware solution is needed

Number of unique users in different hours of the day




Number of Tweets at 00:00
2,860,000
2,850,000
57,000
58,000
10^2
10^1
10^0

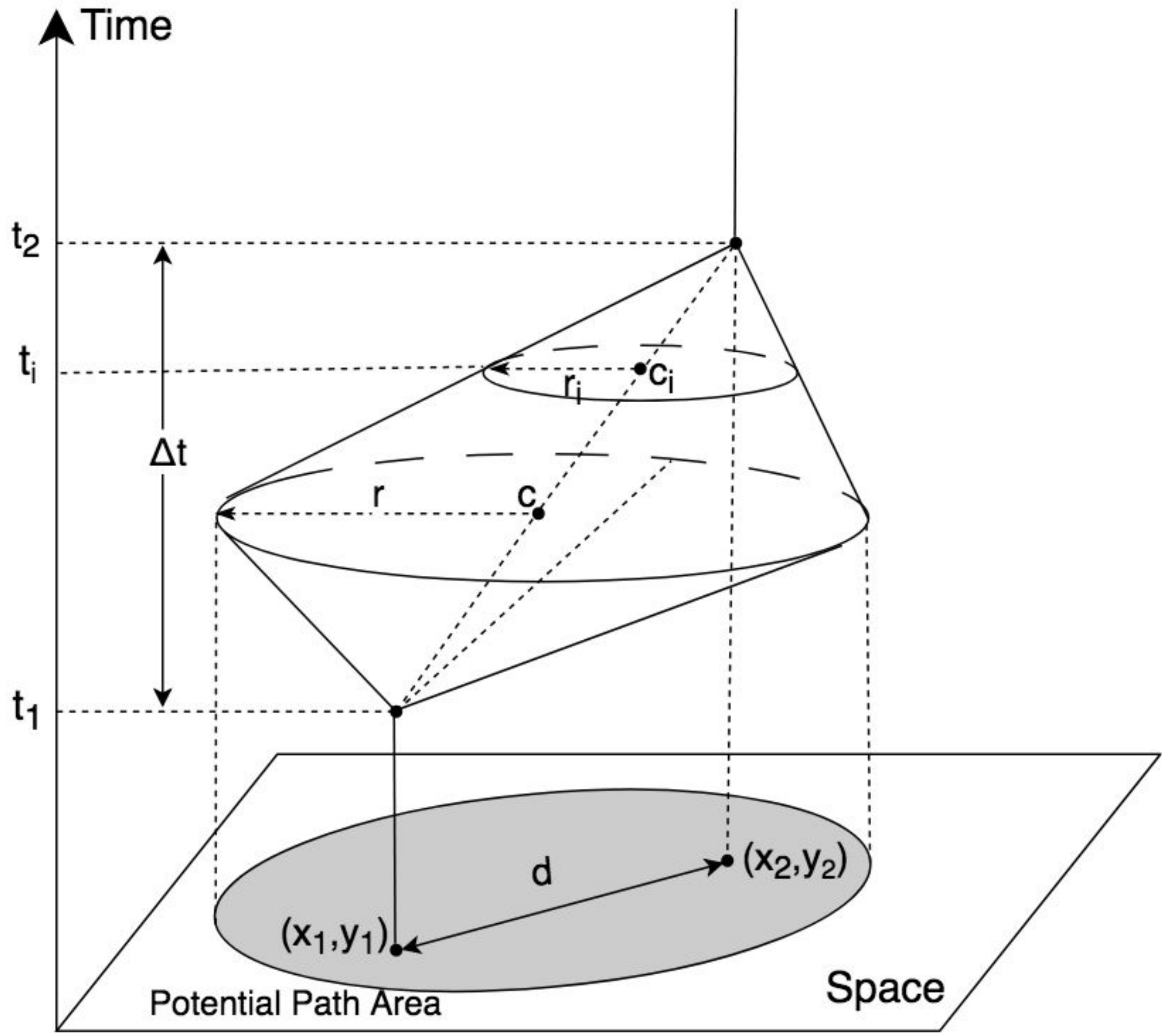

Number of Tweets at 15:00
2,860,000
2,850,000
57,000
58,000
10^2
10^1
10^0


Number of Tweets at 03:00
2,860,000
2,850,000
57,000
58,000
10^2
10^1
10^0
Miami Springs
Virginia Gardens
Miami
Miami Beach
North Bay Village
Coral Gables
West Miami
South Miami
El Portal
Hialeah
Port of Miami

Find the Missing Population

- Location inference based on individual trajectories
  - Given a series of observations [(time, loc), …] of a person, infer the person's location(s) between observations
- The space-time prism (STP)
  - A person's possible activity space between two observation anchors

A STP Diagram of Measuring Activity Space Between Observations



Probability Representation

- 2D Gaussian distribution as the basic units
  - Describe the possible location of any individual at a given moment
    - Center location as the mean value
    - Radius as 2σ (95% confidence range)
  - Mitigate usage bias
    - Frequent users and infrequent users are calibrated to the same temporal scale
  - Accommodate GPS precision
    - Apply a 5m-radius Gaussian distribution on the observation points

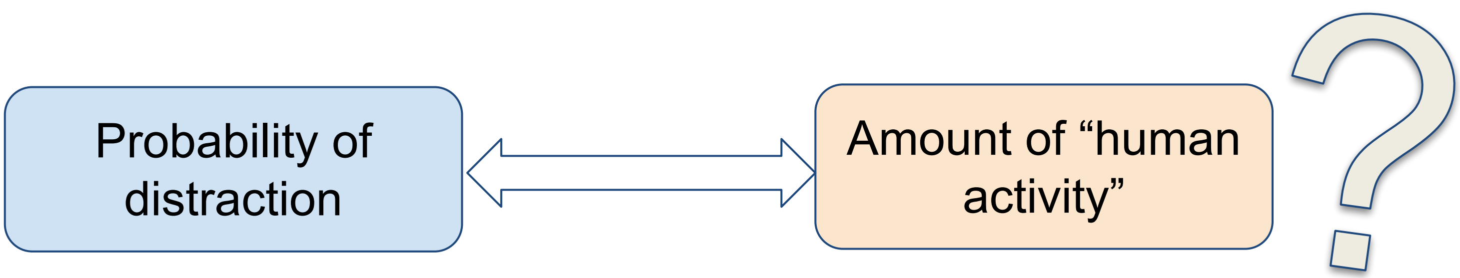
Assumptions of the STP

- Maximum Radius of the STP
  - Upper bound: max speed
  - Lower bound: space/time distance
  - Reasonable estimate: de-facto speed

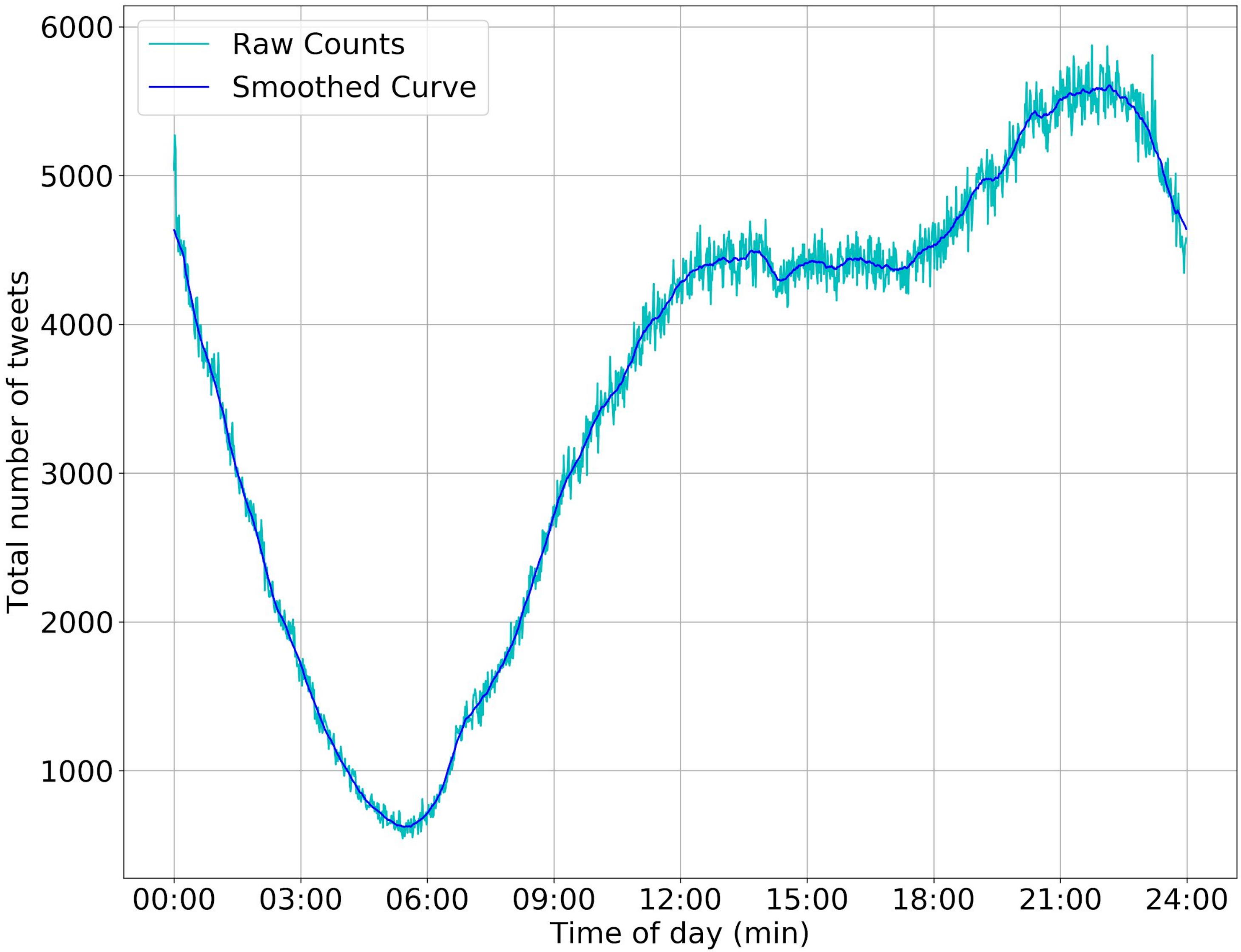
- Validity of the STP
  - A person could be "distracted" between tweets, visiting another place without tweeting
  - The "continuity" between two tweets needs to be measured

Time-based Continuity Estimation

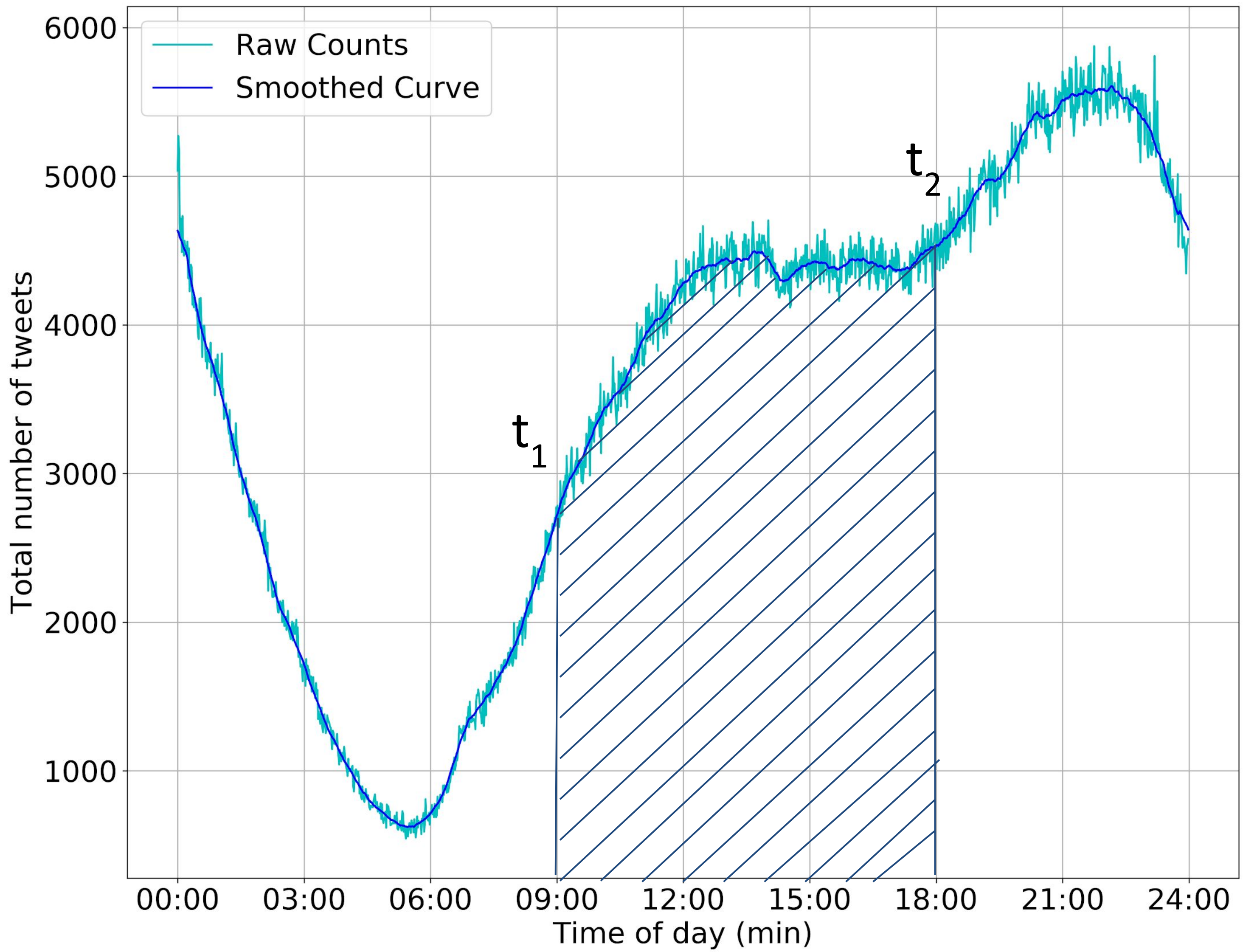
- Consider the time gap between two tweets
  - Gap duration
    - 2 pm - 3 pm V.S. 2 pm - 8 pm
    - 12 pm - 12 pm
  - Gap occasion
    - 12 am - 6 am V.S. 6 am - 12 pm
- Intuition



A temporal activity curve based on the total number of tweets in every minute of the day



Amount of activities as proportion of tweets



Continuity Confidence

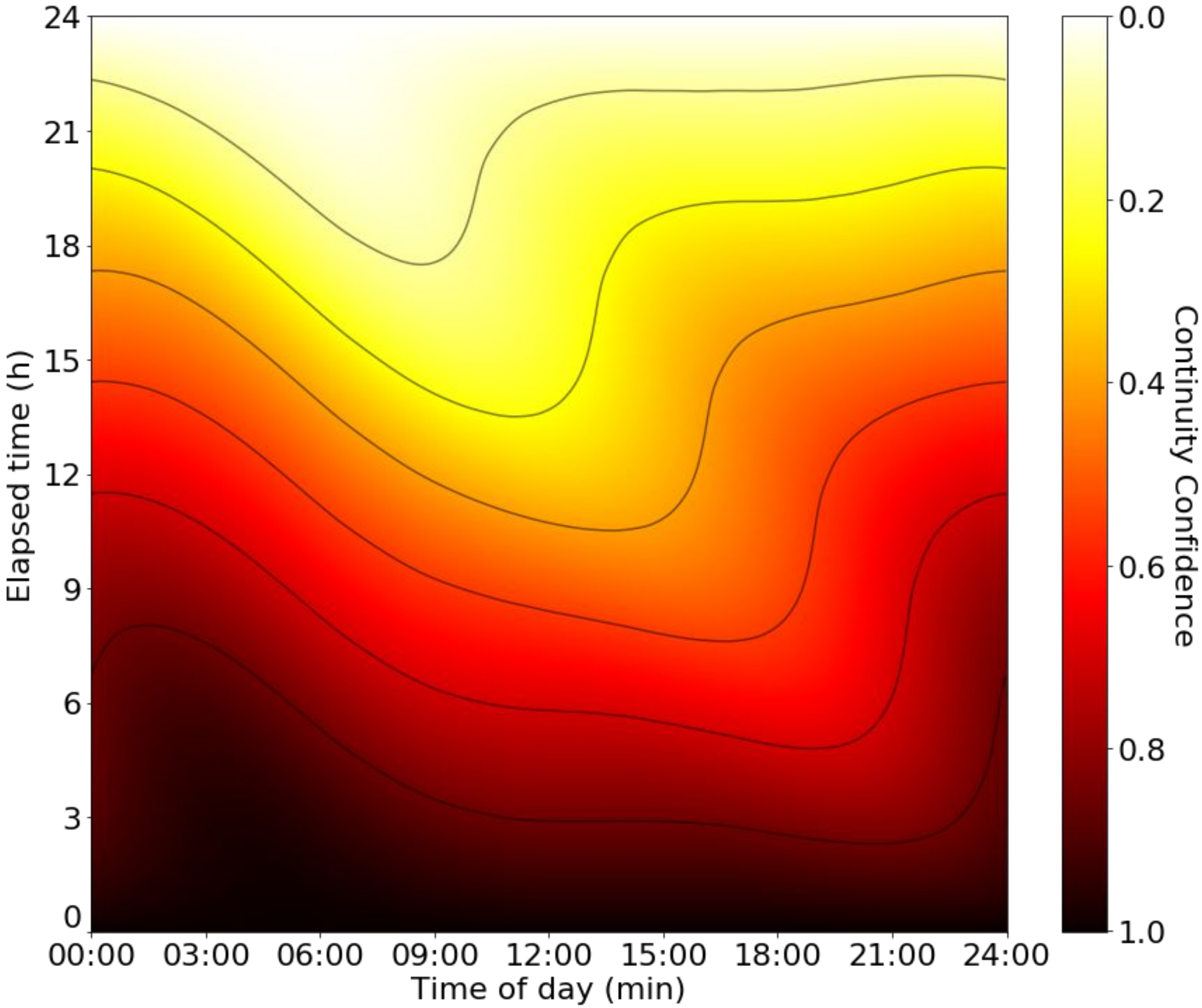
Given the total number of LBSM records as N, and number of records between t_1 and t_2 in the dataset as $N(t_1, t_2)$, the probability of any hidden distraction between t_1 and t_2 is approximated as

$$H(t_1, t_2) \approx \frac{N(t_1,t_2)}{N} \quad .$$

Then, the continuity confidence of the corresponding STP is defined as:

$$C(STP_{t_1,t_2}) = 1 - H(t_1, t_2) \approx 1 - \frac{N(t_1,t_2)}{N} \quad .$$

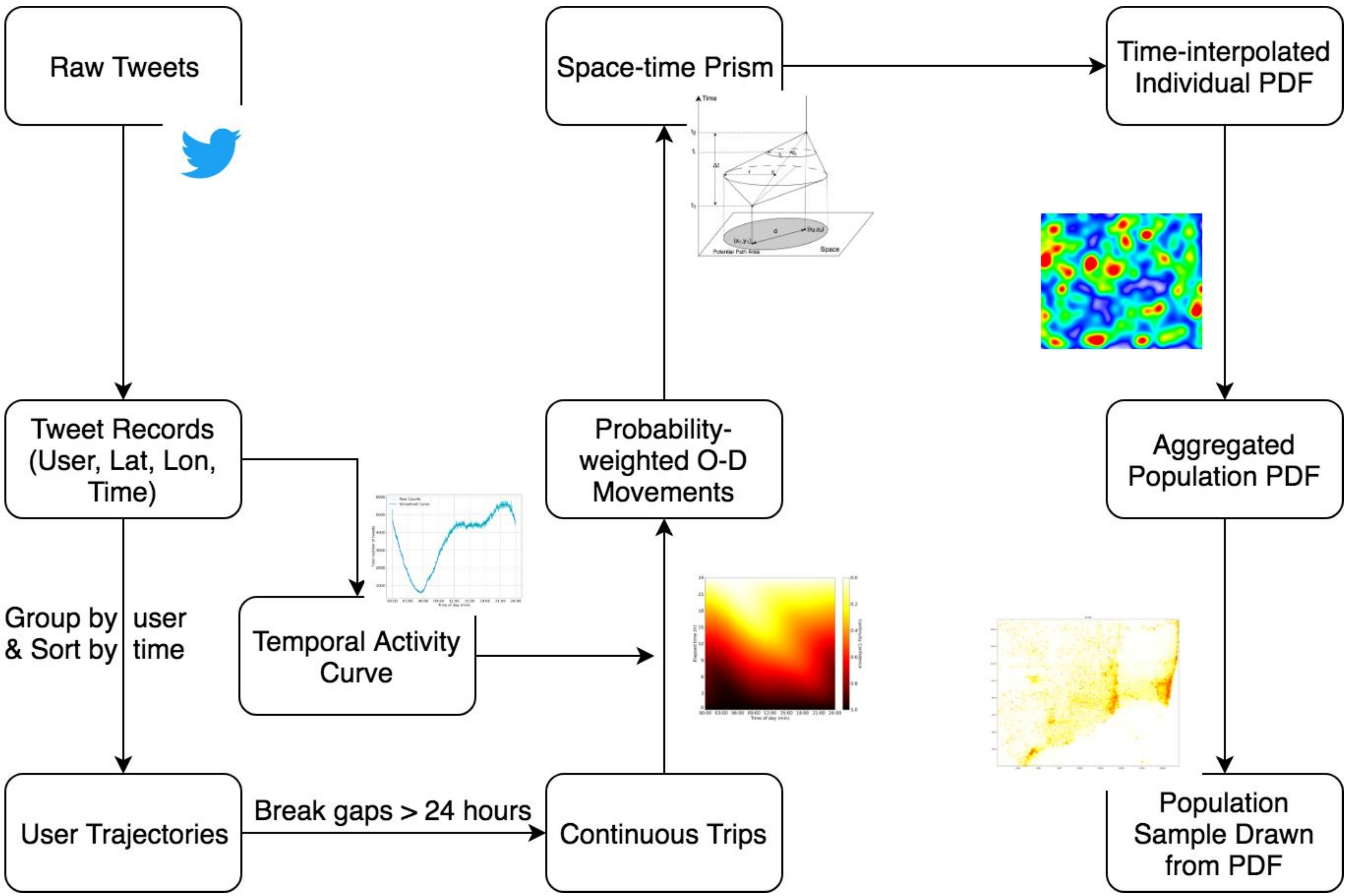
Continuity confidence as a function of the start time and time gap



Conceptual Workflow

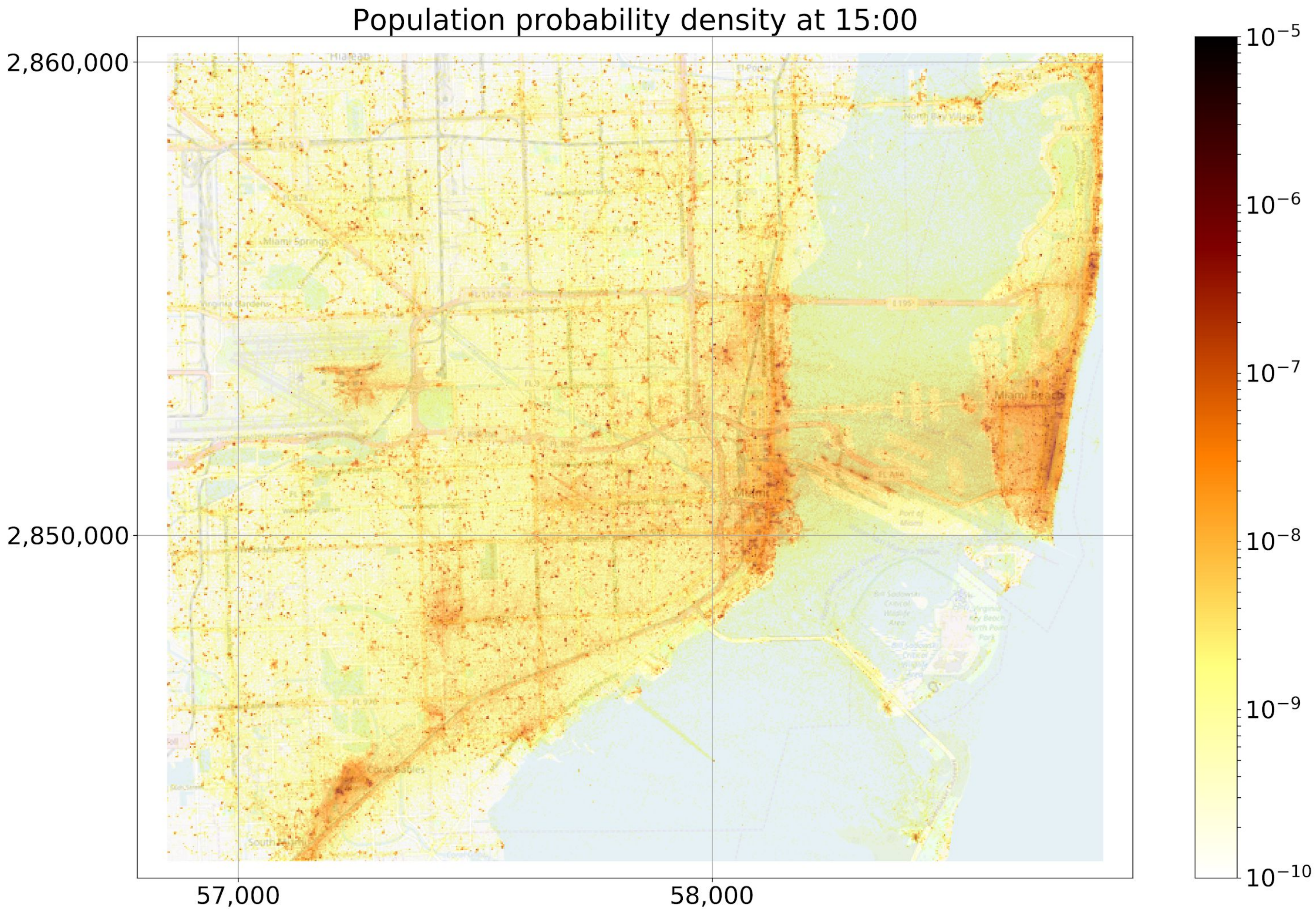
- Obtain user trajectories
- Define continuity confidence
- Build space-time prism
- Interpolate individual Gaussians at every minute
- Aggregate minute-Gaussians to mixed-Gaussians at each hour
- Sum and normalize the whole population distribution

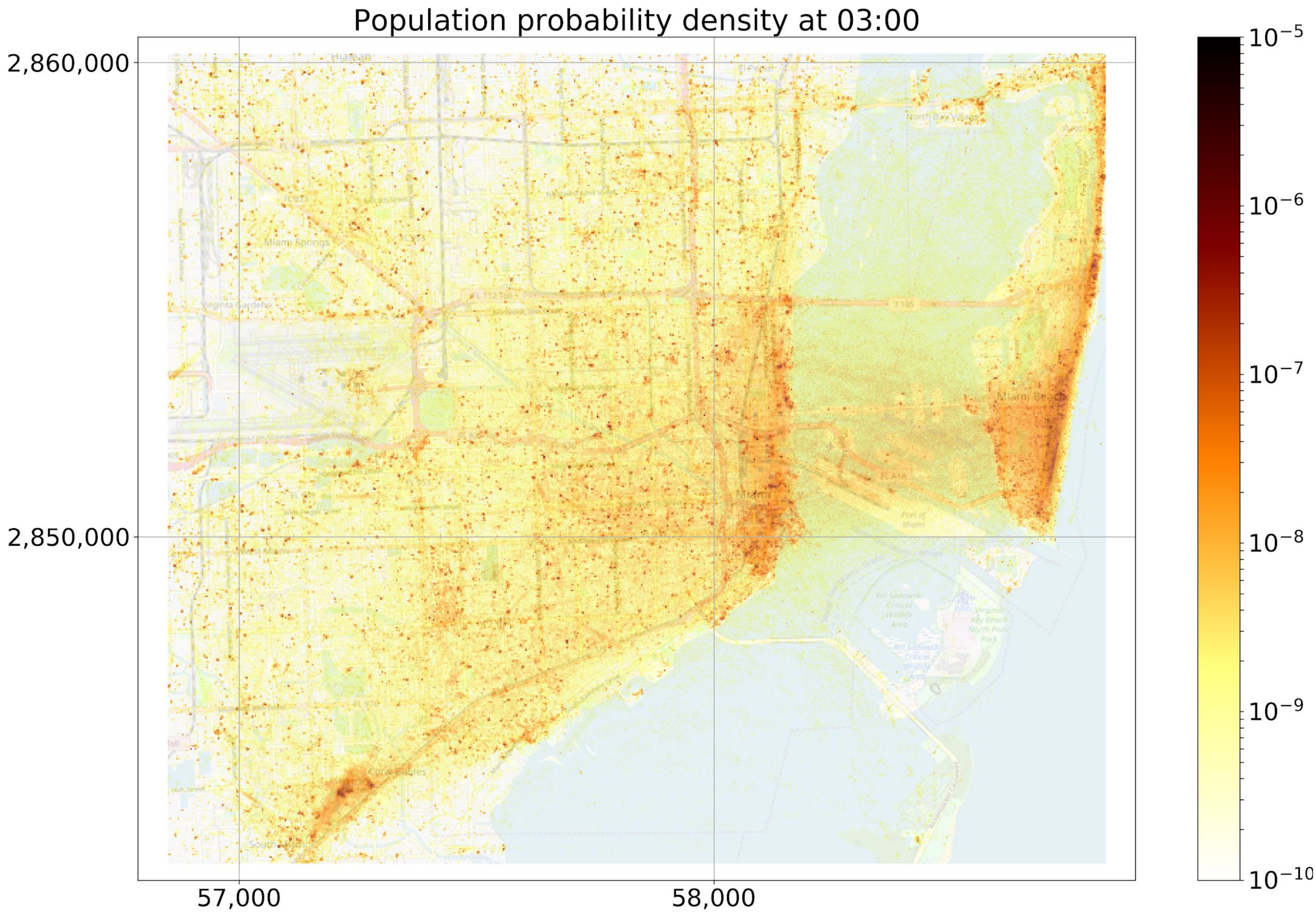
The workflow to draw population samples from raw tweets

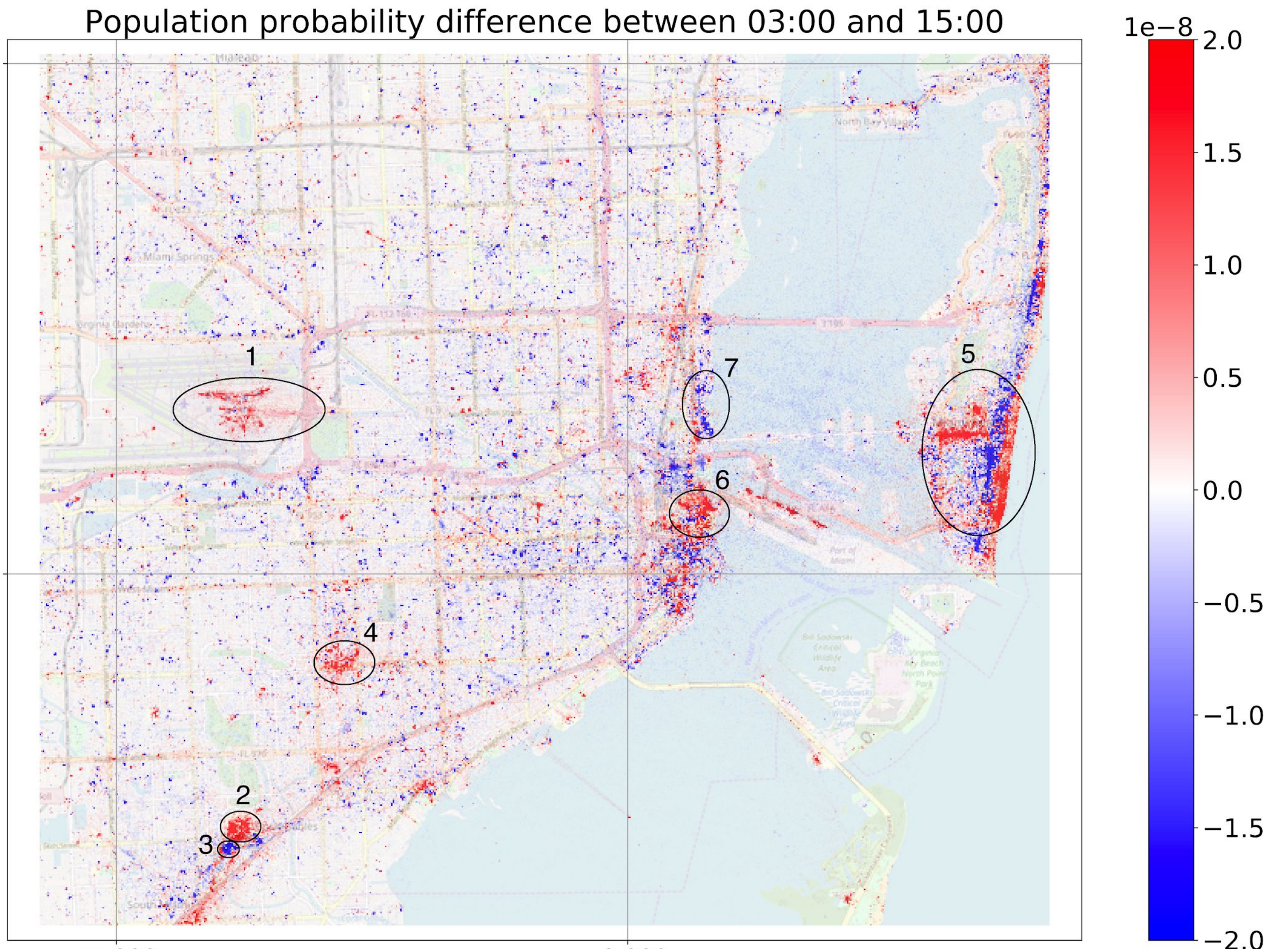


Study Area and Data

- **5.1 million** geo-tagged tweets collected via Twitter API in the city of Miami from Jan. 1st, 2014 to Dec. 31st, 2014
- The specific spatial extent of the study area is 80.119601° W to 80.316665°W, and 25.703935°N to 25.858107° N
- In total, **4.1 million** STPs are constructed, and **424 millions** of Gaussian distribution are interpolated for aggregation
- The final result, **24** probability density distributions are generated for each hour of the day from 0:00 to 23:00, with a spatial resolution of **30x30** square meter


Population probability density at 15:00
2,860,000
2,850,000
57,000
58,000
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
Miami Springs
Miami Beach
Miami


Population probability density at 03:00
2,860,000
2,850,000
57,000
58,000
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
Miami Beach
Miami Springs
Miami



1. Miami International Airport

2. University of Miami

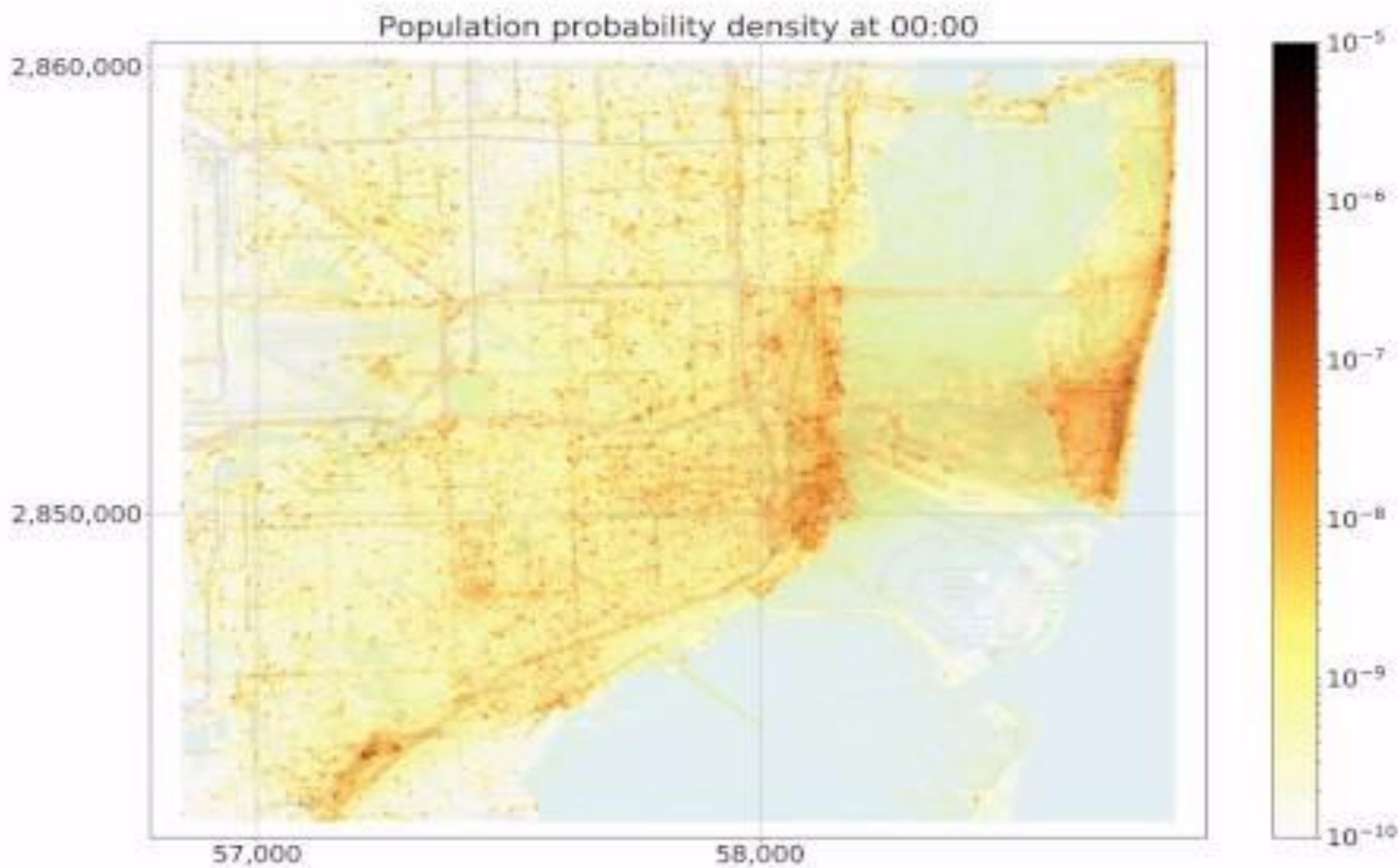
3. Near-campus apartment areas

4. “Miracle Mile” central business district

5. Resorts/hotels and business areas on Miami Beach

6. Downtown Miami

7. Popular hotel clusters near downtown


Population probability density at 00:00
2,860,000
2,850,000
57,000
58,000
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}

Conclusion

- Mitigate some drawbacks of LBSM data
  - Continuity
  - Sparsity
- Demonstrate the feasibility of deriving population distribution at fine spatiotemporal scales

Future Work

- Data refining
  - Advanced methods to filter robots/errors
- Validation
  - With other models or authoritative data
- Parameter calibration
  - Seasonal dynamics
  - Weekdays v.s. weekends
- Applications
  - Integration with census data
  - Input to agent-based models
  - Temporal activity curve and place type

Acknowledgements

- National Science Foundation
  - 1443080
  - 1743184

Thanks!

- **Comments/Questions?**
- **Email:** dyin4@illinois.edu