Approximate Computing

An effective likelihood-free method with statistical guarantees

Ryan Gross

Suzanne Thornton

Dr. Minge Xie

DIMACS REU 2018

Madison Square Park data set

Data

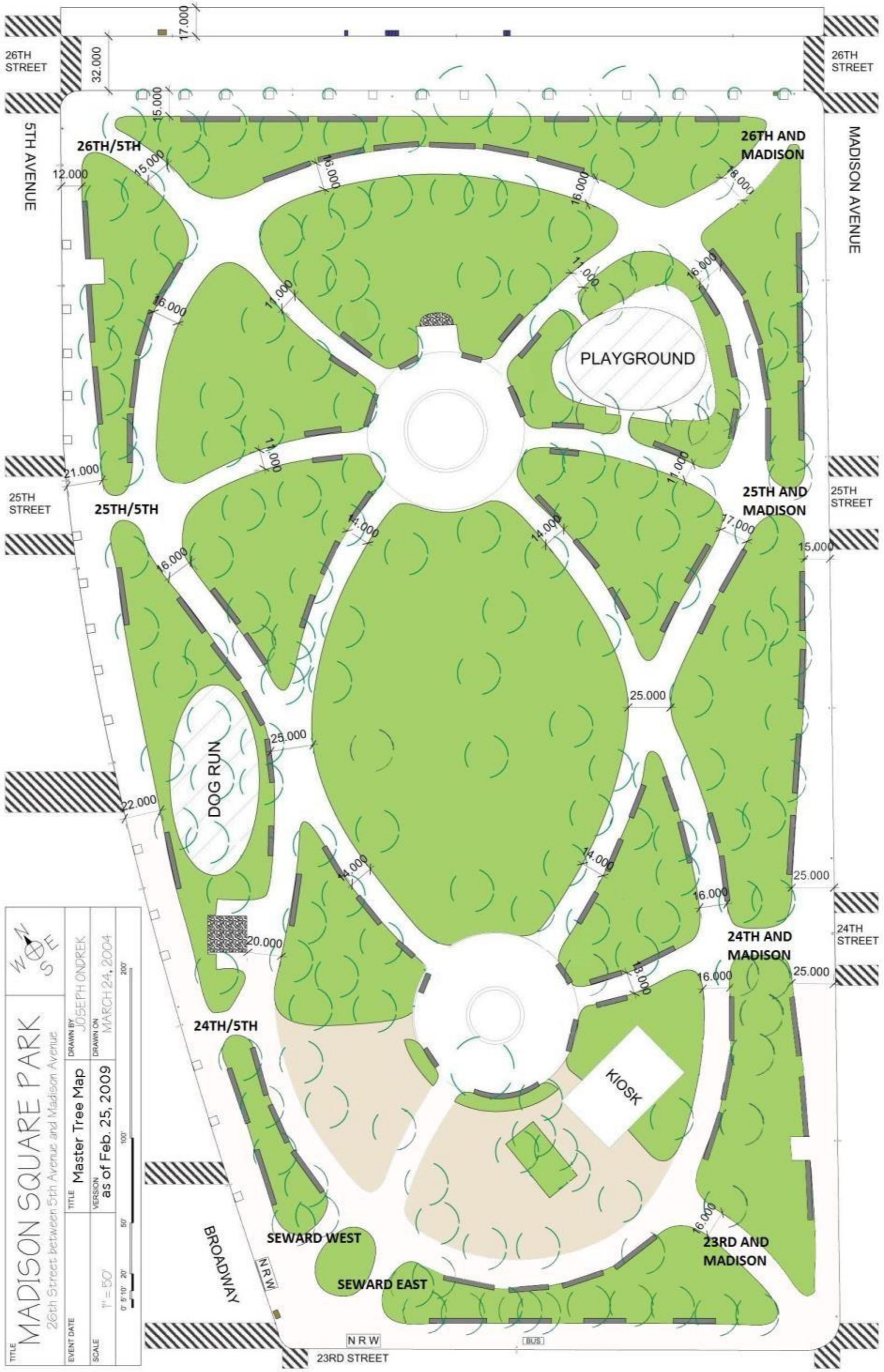
- 2015-2017 entry/exit counts per day per location

The Problem

- 2 of 9 entrances equipped with counters at a time
- Lost data from “transition days”
- Time of day, weather patterns, events not accounted for

Goal

- Estimate total number of park users over a given time period



The Process

Data Cleanup/ Investigation
- Initial data analysis
- Determine summary stats

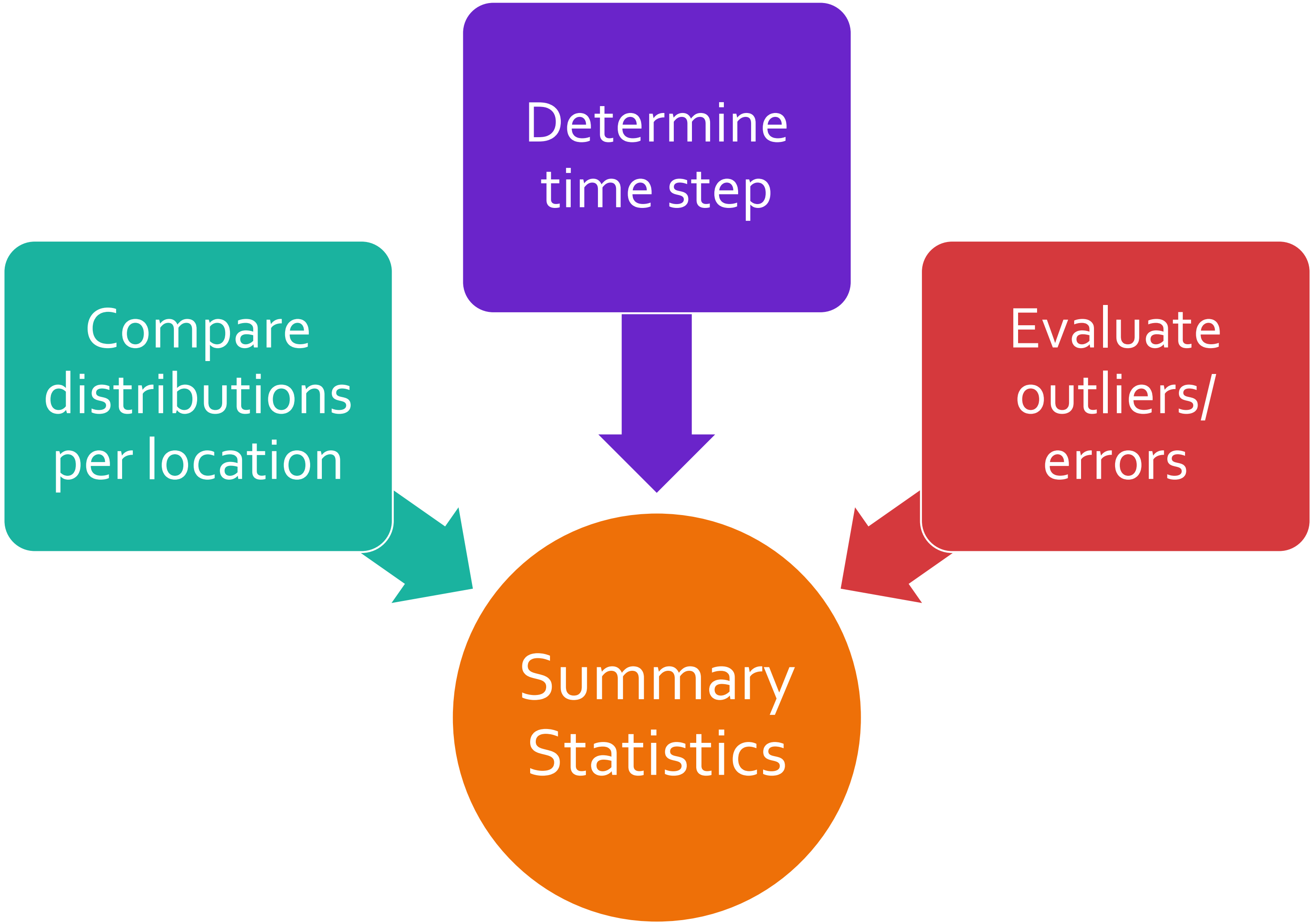
Model
- Poisson point process, stochastic matrix
- Multiscale (micro-macro)

Run simulations
- Replicate missingness
- Determine simulated counts

Approximate Computing
- ABC, ACC methods
- Compare simulation to data

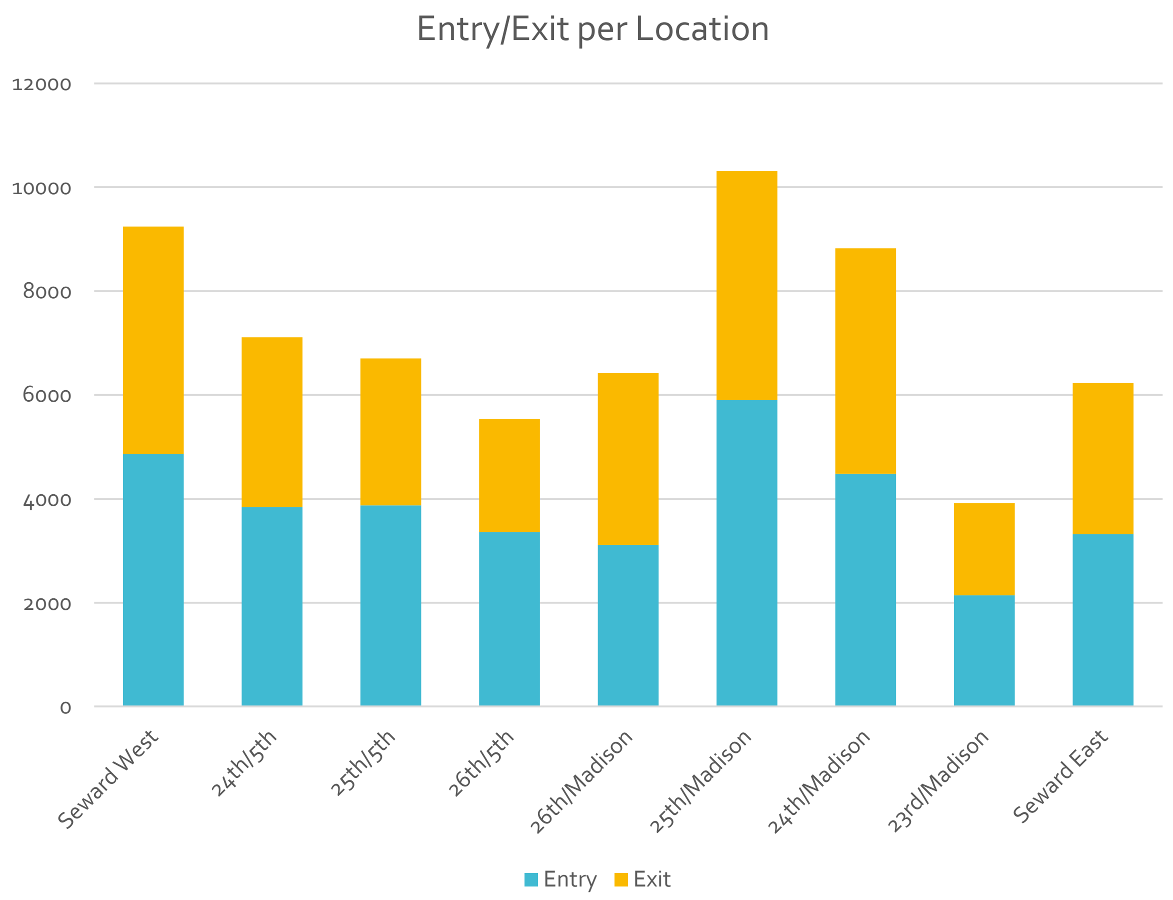
Refine and repeat
- Adjust model
- Update summary stats

Initial Data Analysis

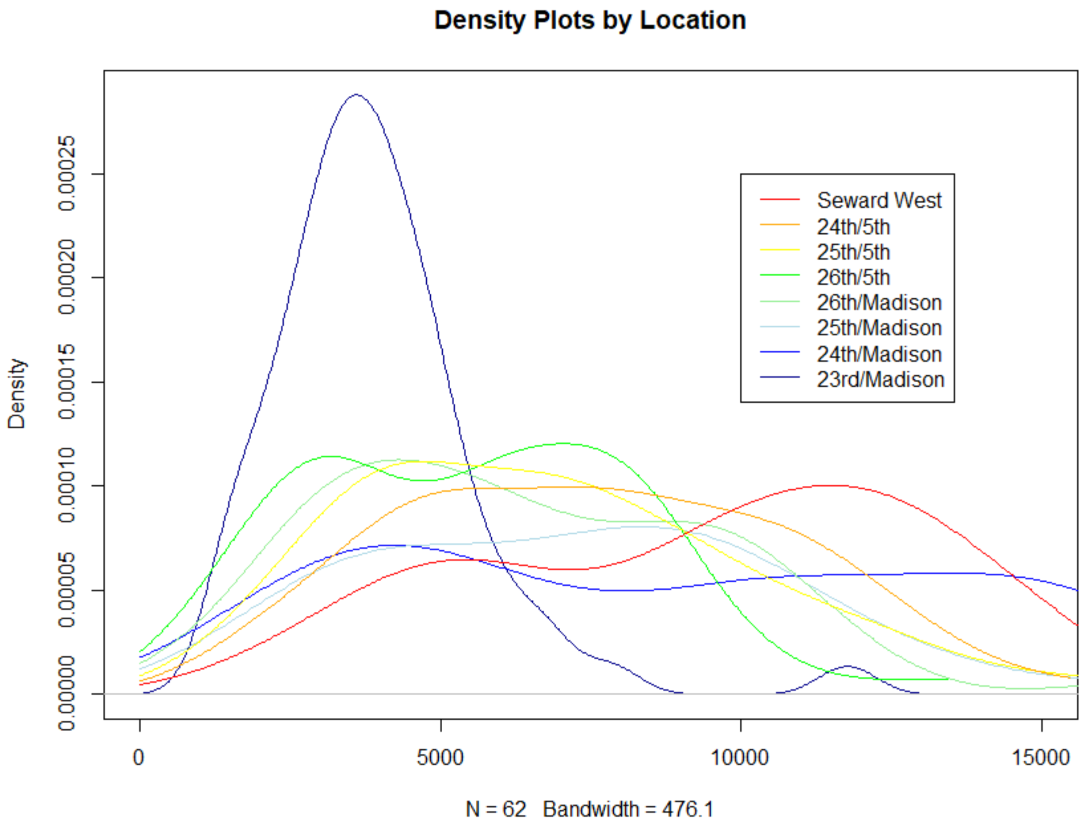


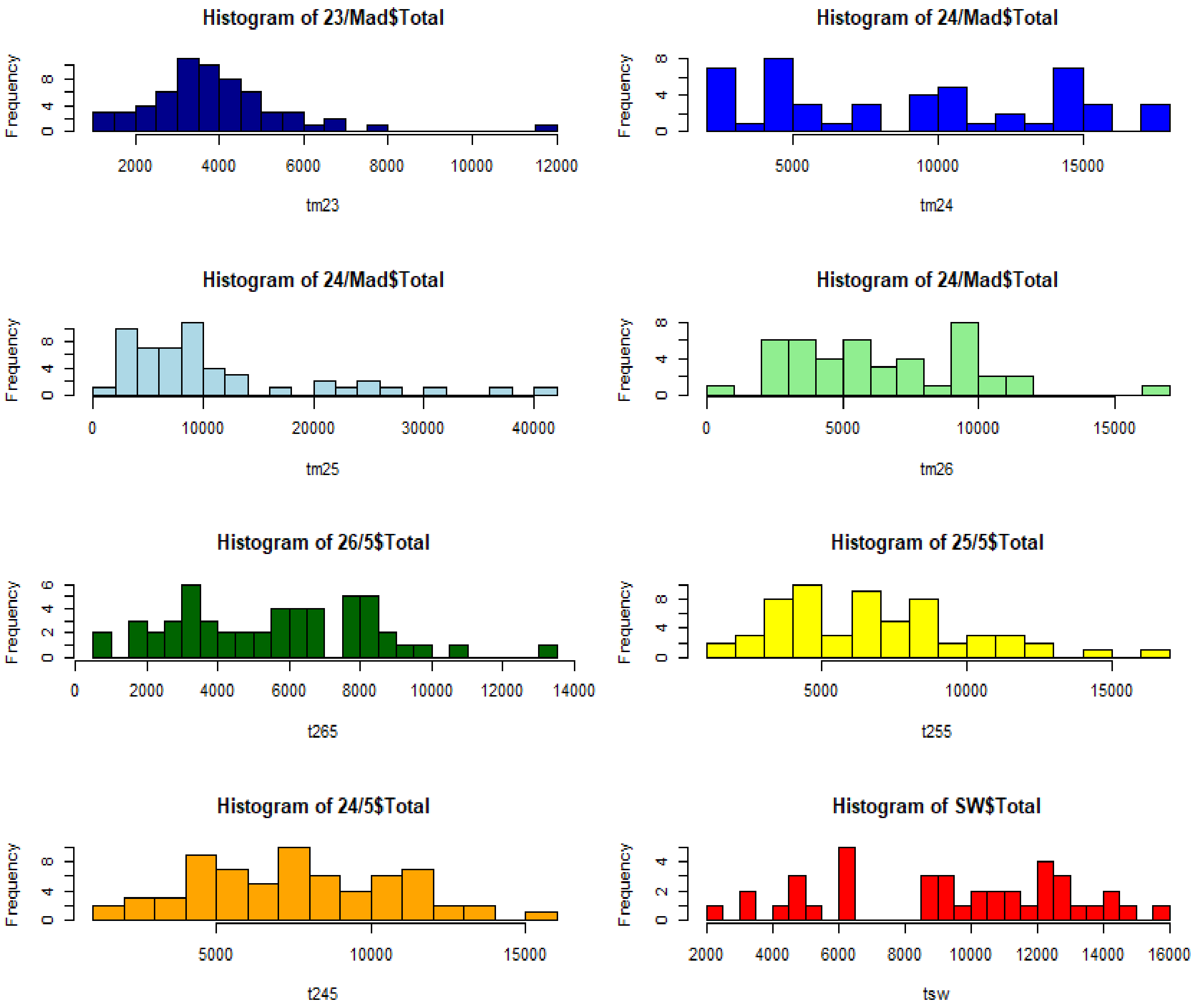
Total Counts

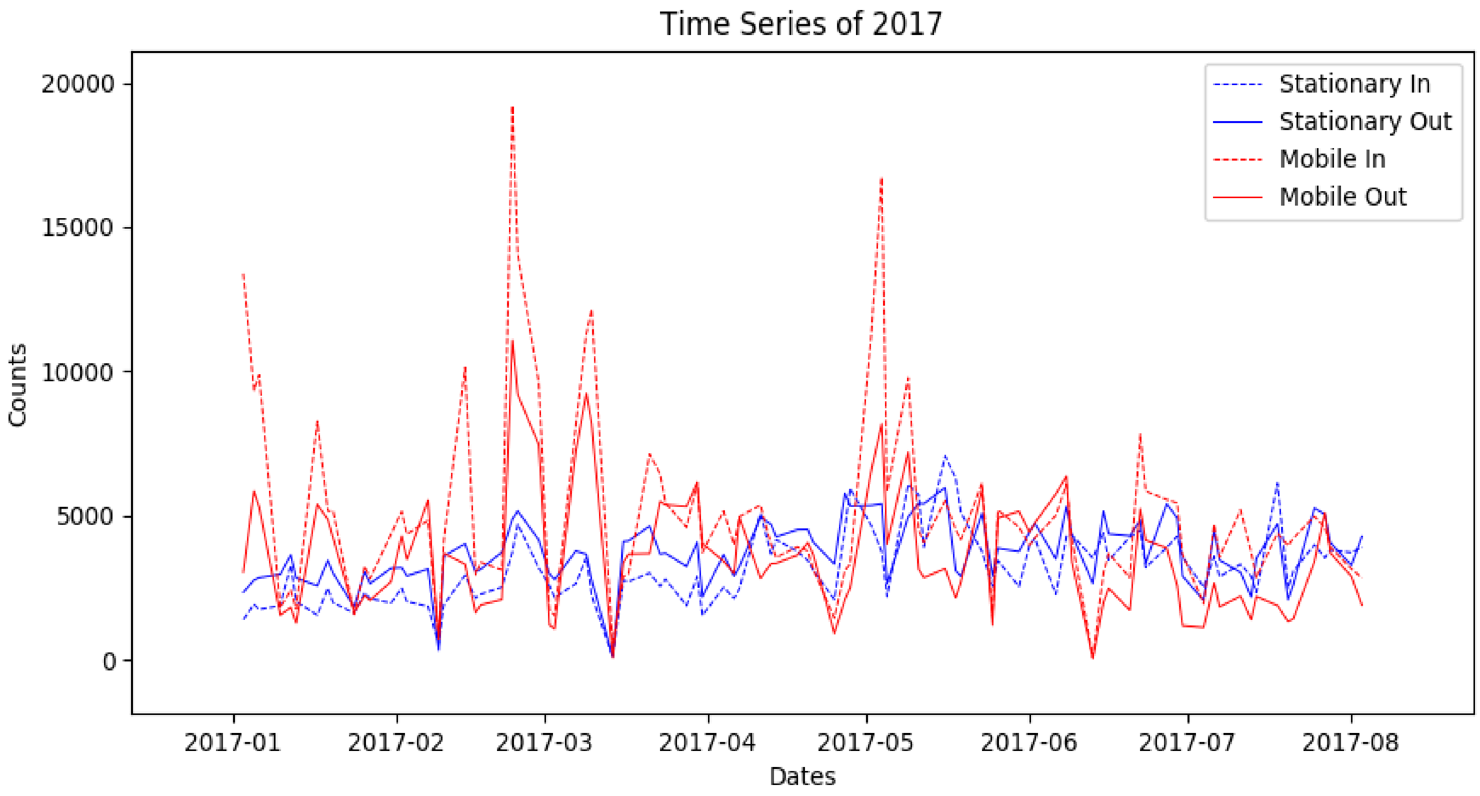
Locations	Avg. Daily Entry	Avg. Daily Exit	Avg. Daily Total
Seward West	4868	4376	9088
24th/5th	3846	3267	7113
25th/5th	3871	2831	6702
26th/5th	3365	2175	5540
26th/Madison	3116	3304	6420
25th/Madison	5905	4403	10307
24th/Madison	4484	4339	8823
23rd/Madison	2143	1772	3915
Seward East	3321	2907	6228



Distributions/Densities





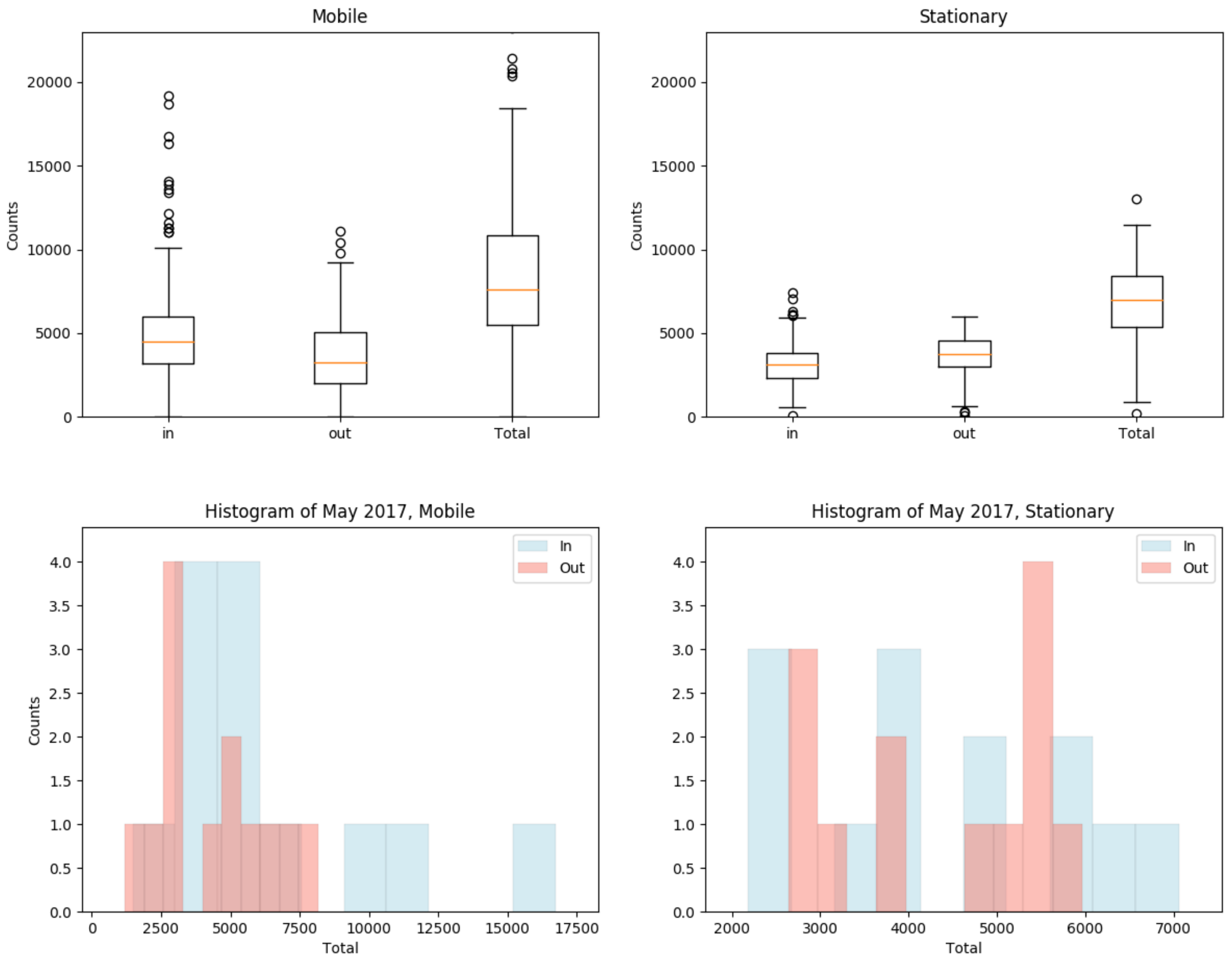


Time Step

- Use weekly data
  - Daily: too much variation
  - Monthly: too few observations

Counter Errors

- Entry counts generally higher than exit counts
  - All locations, periods of time
  - Largely due to outliers
- Must account for in simulations
  - Separate entry/exit data in comparisons

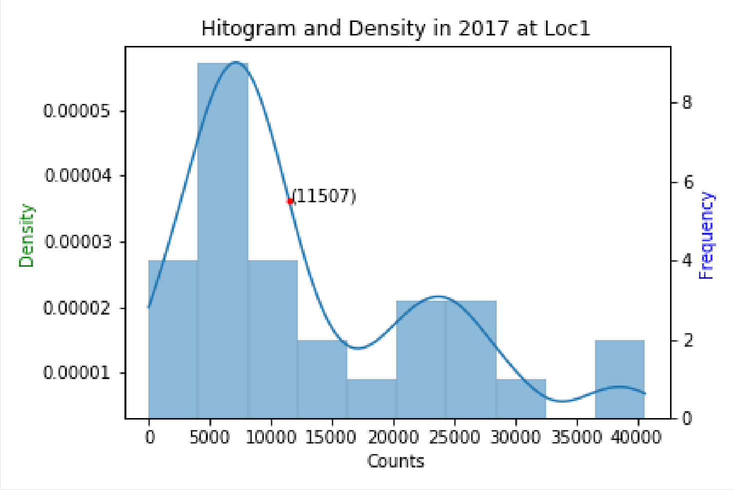


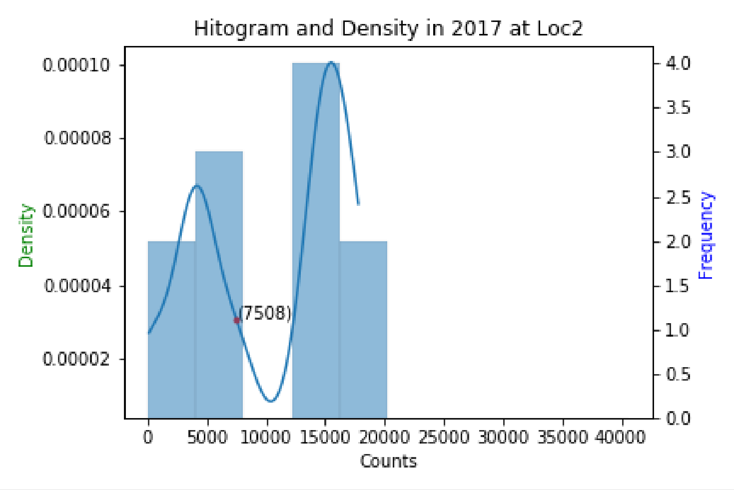
Summary Statistics

- Relative rate of entry/exit (weekly basis)
- 18 parameters total
- Used in ACC method
- Later: look to decrease number of parameters
  - Go back to initial analysis to find relationships

Locations	Rate In	Rate Out
Seward West	0.14	0.15
24th/5th	0.11	0.11
25th/5th	0.11	0.10
26th/5th	0.10	0.07
26th/Madison	0.09	0.11
25th/Madison	0.17	0.15
24th/Madison	0.13	0.15
23rd/Madison	0.06	0.06
Seward East	0.10	0.10

Simple Model Example





- Time period: monthly (April-June)
- Outliers/errors: ignored
- Summary statistic: mean
- Estimation method: Gaussian kernel

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

	Counts in 2017	Counts in May, 2017	Total Average in May, 2017	Total Average from April to June, 2017
Transition	49	10	--	--
Loc1	15	2	17360	11507
Loc2	9	1	17810	7508
Loc3	14	0	--	5398
Loc4	7	1	16970	11333
Loc5	11	2	7260	7613
Loc6	9	2	9208	8822
Loc7	12	4	6571	6571
Loc8	3	1	12032	12032

Sum

39599 visitors in May

Modeling

Stochastic model, Spatial point process, Multiscale model

Stochastic Matrix

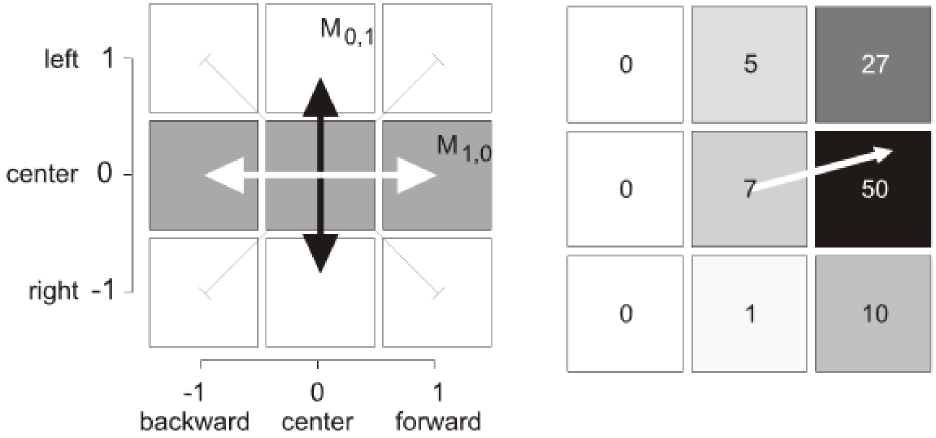
- Based on square grid structure and uses 3x3 transition matrix for individual movements
- Updates probabilities based on status of surrounding cells (obstacles, other pedestrians)

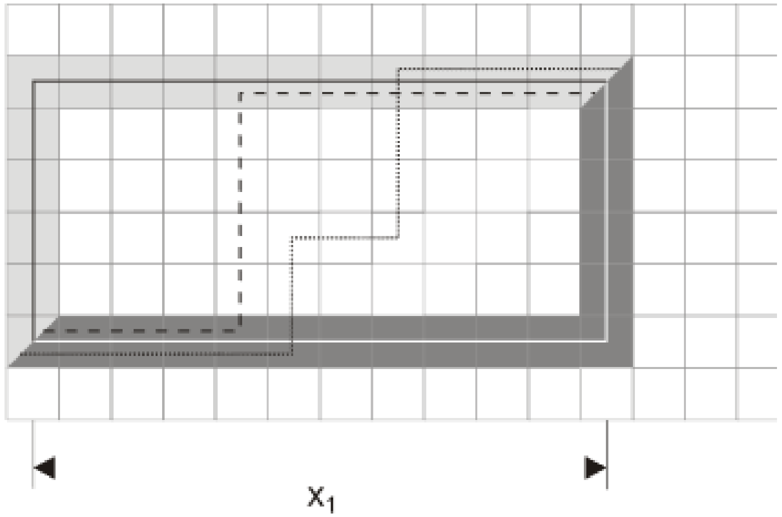
$$p_{\text{forward or left}} = \frac{1}{2}\ \sigma^2 + \mu^2 + \mu$$

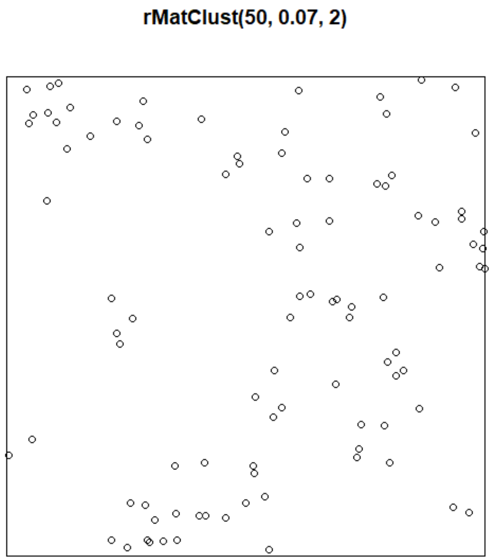
$$p_{\text{center}} = 1 - \ \sigma^2 + \mu^2$$

$$p_{\text{backward or right}} = \frac{1}{2}\ \sigma^2 + \mu^2 - \mu$$

- Issues:
  - Only a microscopic look at individual pedestrians
  - No data to determine parameters for the model







Spatial Poisson point process

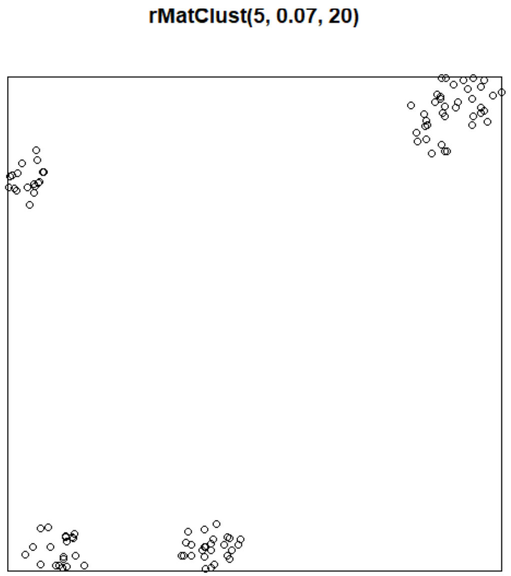
- Spatial point process is a random pattern of points (e.g. pedestrians) in d-dimensional space

$$N_t = \text{number of points arriving up to time } t$$
$$= \sum_{i=1}^{\infty} \mathbf{1}\{T_i \leq t\},$$

- Poisson distribution: models arrival times

$$P\{N(B) = n\} = \frac{(\Lambda(B))^n}{n!} e^{-\Lambda(B)}$$

- Issues:
  - Macroscopic view does not account for movement within the park
  - Too simplified for our park data (many parameters, missing data)



Multiscale (micro-macro) model

- Microscopic model:
  - Tracks pedestrians individually
  - System of ordinary differential equations
  - Discrete

$$\begin{cases} \dot{X}^k(t) = V^k(t) \\ \dot{V}^k(t) = F^k(t, X^1, \ldots, X^N, V^1, \ldots, V^N), \end{cases} \qquad k = 1, \ldots, N.$$

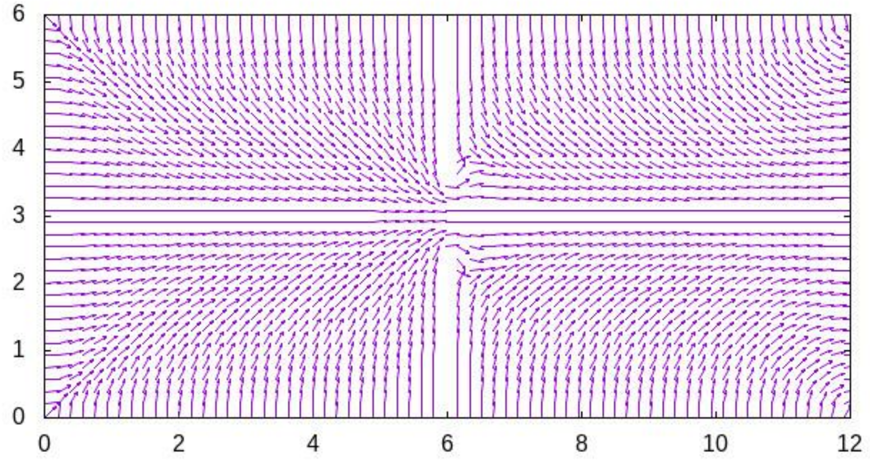
- Macroscopic model:
  - Tracks crowd density
  - Velocity vector field
  - Continuous

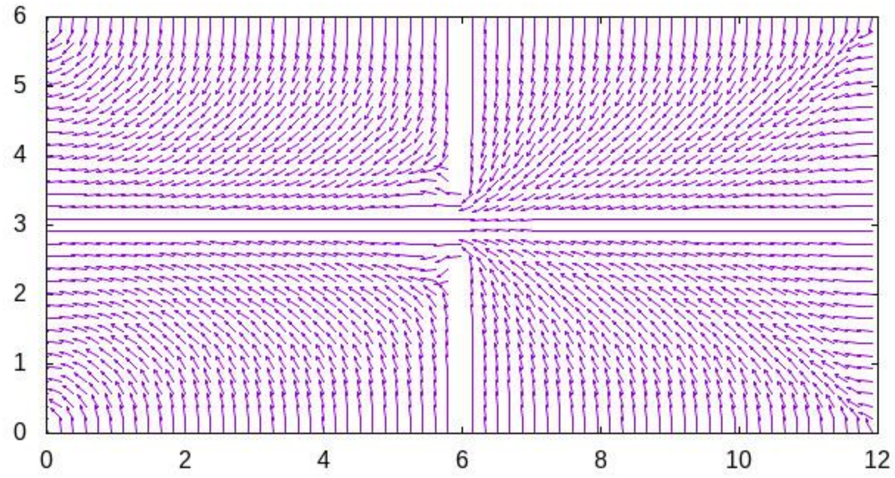
$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0, & t > 0, \quad x \in \Omega \\ \dfrac{\partial v}{\partial t} + (v \cdot \nabla) v = a(\rho, v), & t > 0, \quad x \in \Omega \end{cases}$$

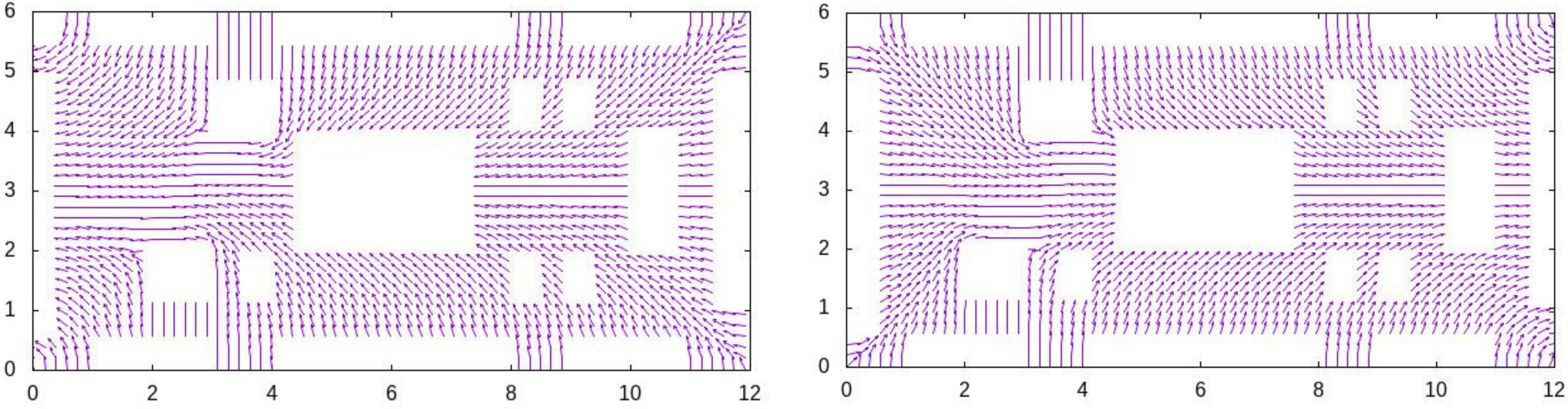
- Key: combine them into computational algorithm
  - We have the code!

The Code

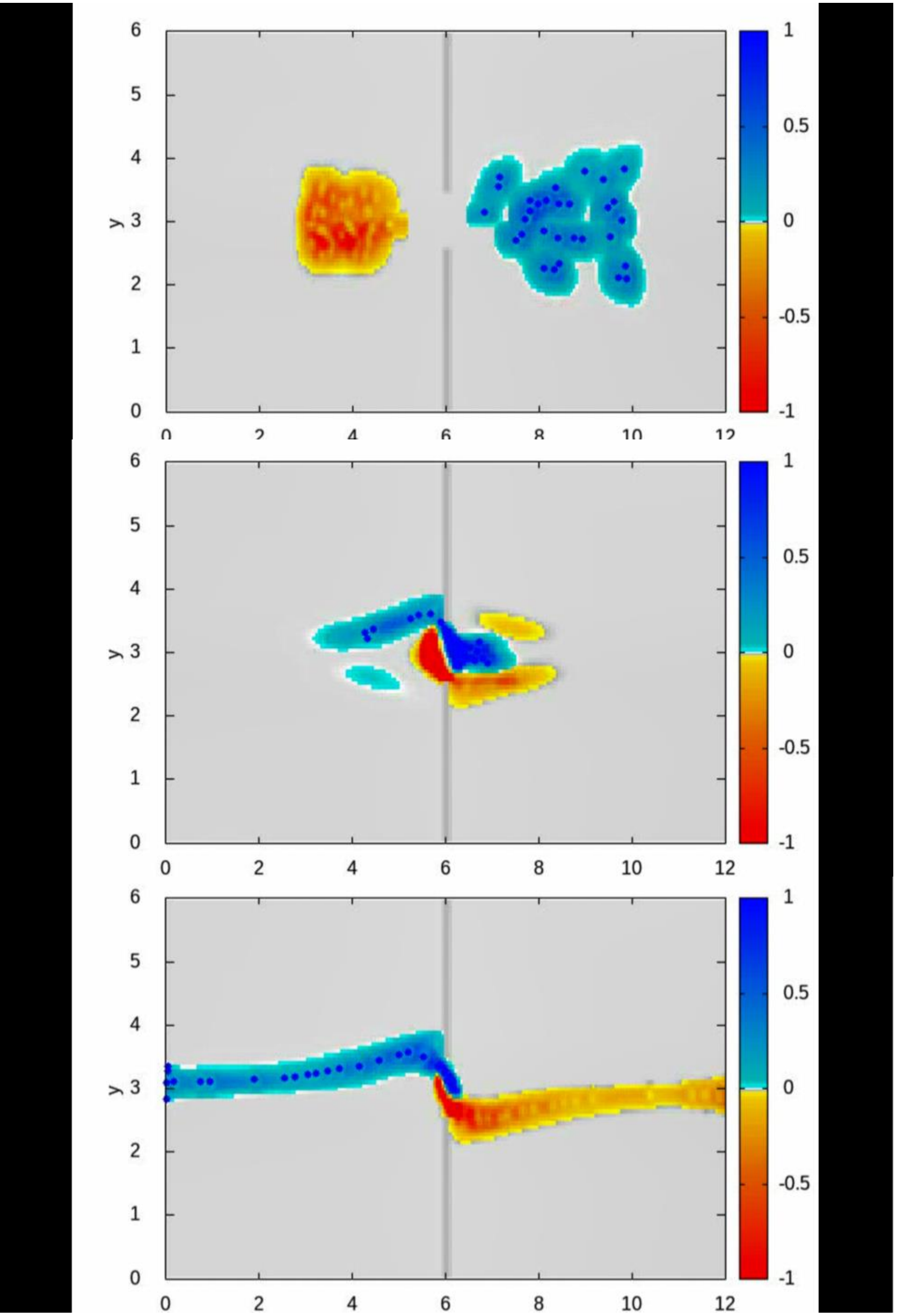
- Source: Dr. Piccoli of Rutgers—Camden
  - Models traffic flow using PDEs
- Produces velocity vector field
  - Individual vectors correspond to pedestrian movement
  - Two populations model inward and outward flow







Applying multiscale model to Madison Square Park



Simulations

Still to come:

- Adjust rates of inflow/outflow by entrance
- Run the simulations
- Calculate counts
- ABC/ACC method

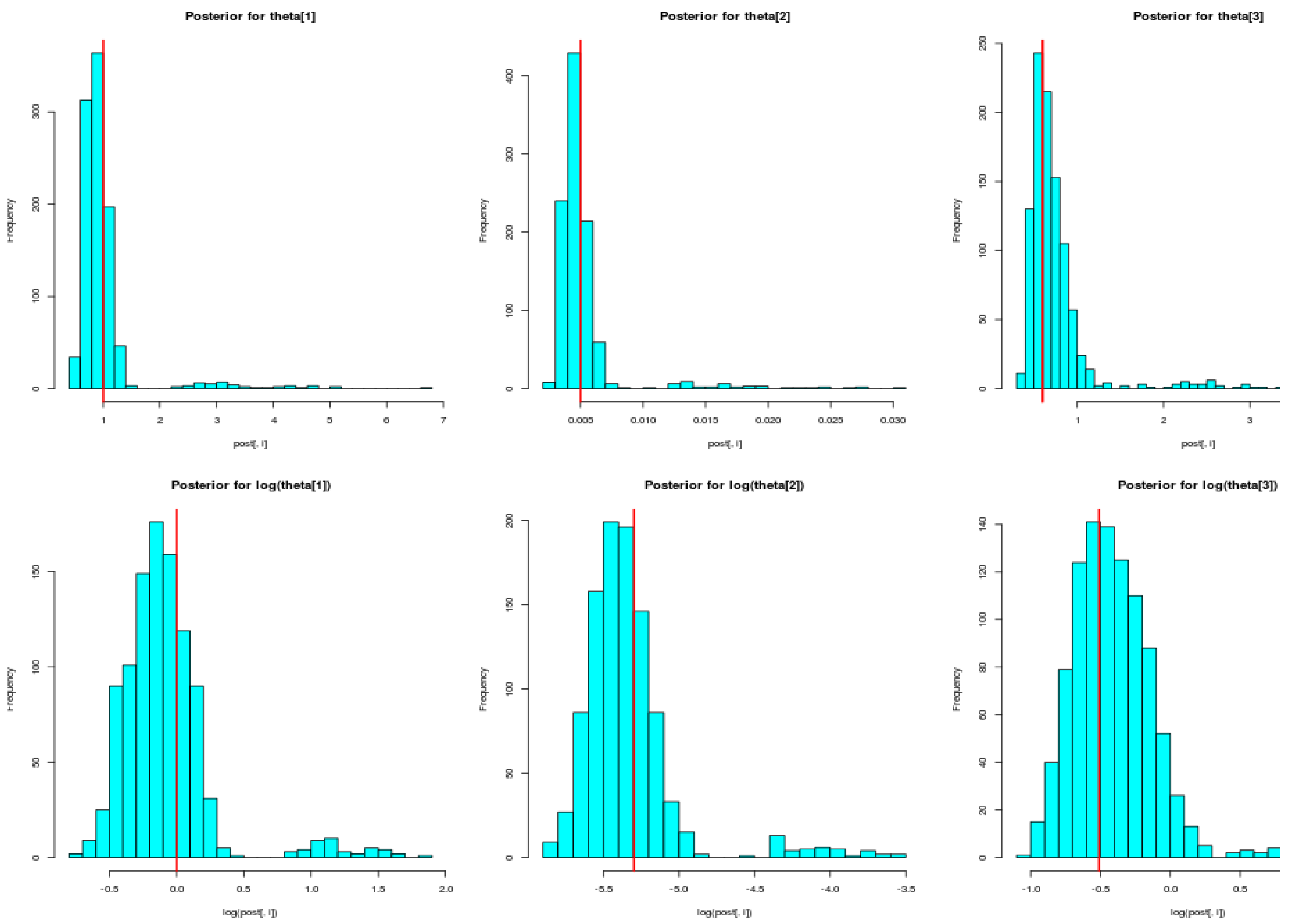
Approximate Computing: ABC method

The algorithm:

1. Generate $\theta_1,\ldots,\theta_N \sim \pi(\theta)$;
   - Use prior assumptions to come up with a set of parameters

2. For each i, simulate $x^{(i)} = \{x_1^{(i)},\ldots,x_N^{(i)}\}$ from M_θ;
   - Now using the model, run simulations to produce a simulated dataset

3. For each i, accept θ_i if $\rho(S^*_n, s_{obs}) \le \varepsilon_n$
   - Compare simulated data to observed data (using summary stats). If they are close enough, accept θ_i. If not, discard θ_i and return to step 1



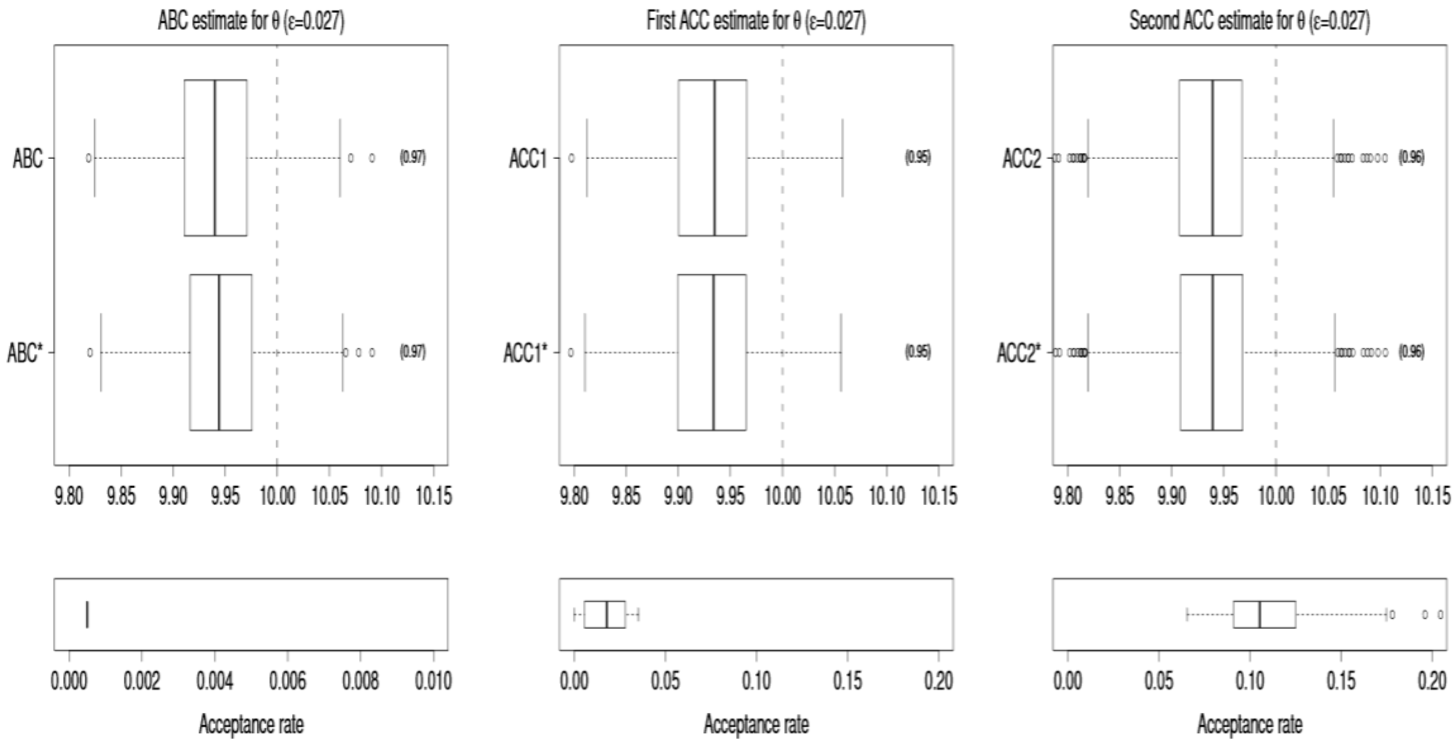
Approximate Computing: ACC method

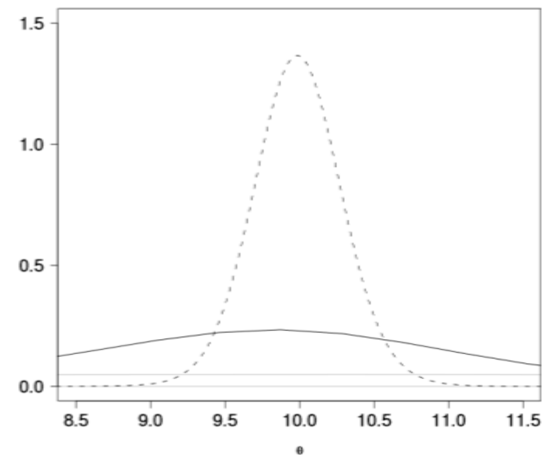
The algorithm:

1. Generate $\theta_1,\ldots,\theta_N \sim r_n(\theta)$;
   - Instead of prior assumption, free to select data-dependent distribution r_n from which parameters are generated

2. and 3. identical to ABC method

- Key: Data-dependent ACC has computational advantage over ABC





References

- Thornton, S., & Xie, M. (2018). Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees. arXiv:1705.10347
- Cristiani, E., Piccoli, B., & Tosin, A. (2014). *Multiscale Modeling of Pedestrian Dynamics* (Vol. 12). Springer International Publishing Switzerland. doi:10.1007/978-3-319-06620-2
- Cristiani, E., Piccoli, B., & Tosin, A. (2011). Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Modeling & Simulation*, 9(1), 155-182. doi:10.1137/100797515