Statistical approaches to mobility modeling

Abel Rodríguez - UC, Santa Cruz

ATD Workshop
September 14, 2017

Outline

- Previous work: Socially- and Geography-Aware (SAGA) Mobility Modeling.
  - Motivation.
  - Model.
  - Results.
- Proposed work: Point process models for mobility
  - Stylized features of human behavior.
  - Models for position (population-based models).
  - Models for movement (individual-based models).

Motivation for SAGA: Evaluating networking protocols

- The use of discrete-event simulation (e.g., ns-3) to evaluate networking protocols and equipment is well stablished.
- A number of infrastructure-less network architectures exist: MANETs, WSNs, DTNs.
- Evaluation of protocols requires that we have an understanding of how nodes move, but collecting real GPS traces can be a complex process fraught with privacy issues.
- Some real traces are available, but they are limited and do not allow us to investigate the variability in performance.
  - Construct mobility simulators with the right qualitative features (non-uniform spatial density, preferential attachment).
  - Need to "calibrate" the simulator to reflect real human behavior.

Our agent-based model: Basic rules

- Nodes are modeled independently from each other.
- Nodes select a destination, move at a constant speed towards it, and then stop at that location for some time before repeating.
- The destination, speed of movement and pause time are selected according to appropriate probability distributions that are independent of the current location.
  - Not realistic, but necessary given sparsity of available data.
- Intensity function is approximated using a step function.
  - The region of interest is partitioned into a a series of non-overlapping cells, usually (but not necessarily) according to a regular grid.
  - Intensity is assumed to be uniform within the cell.

Agent-based model: Incorporating social interactions

- We assume that nodes can be grouped into "communities" with similar behaviors:
- The baseline spatial intensity as well as the speed and pause time distributions can in principle differ across communities.
- In addition to the a community specific baseline, the attraction of an individual to a given cell depends on the number of individuals of each community that are currently located in it.
  - Community here needs to be interpreted in a more general sense than "social community" $\Rightarrow$ Members of a community are not necessarily characterized by often being together.

Calibrating the model

- Community identification using clustering. Node features include:
  - Number of times each of the l cell was visited.
  - The average speed of the individual and its standard deviation.
  - Average pause time and its standard deviation.
- Multinomial logit model for transition probabilities:

$$\log\left\{\frac{\Pr(y_{n,t}=i)}{\Pr(y_{n,t}=1)}\right\} = c_{i,\xi_n} + \sum_{l=1}^{K} \theta_{\min\{\xi_n,l\},\max\{\xi_n,l\}} \log\left(1 + a_{i,l,t-1}\right)$$

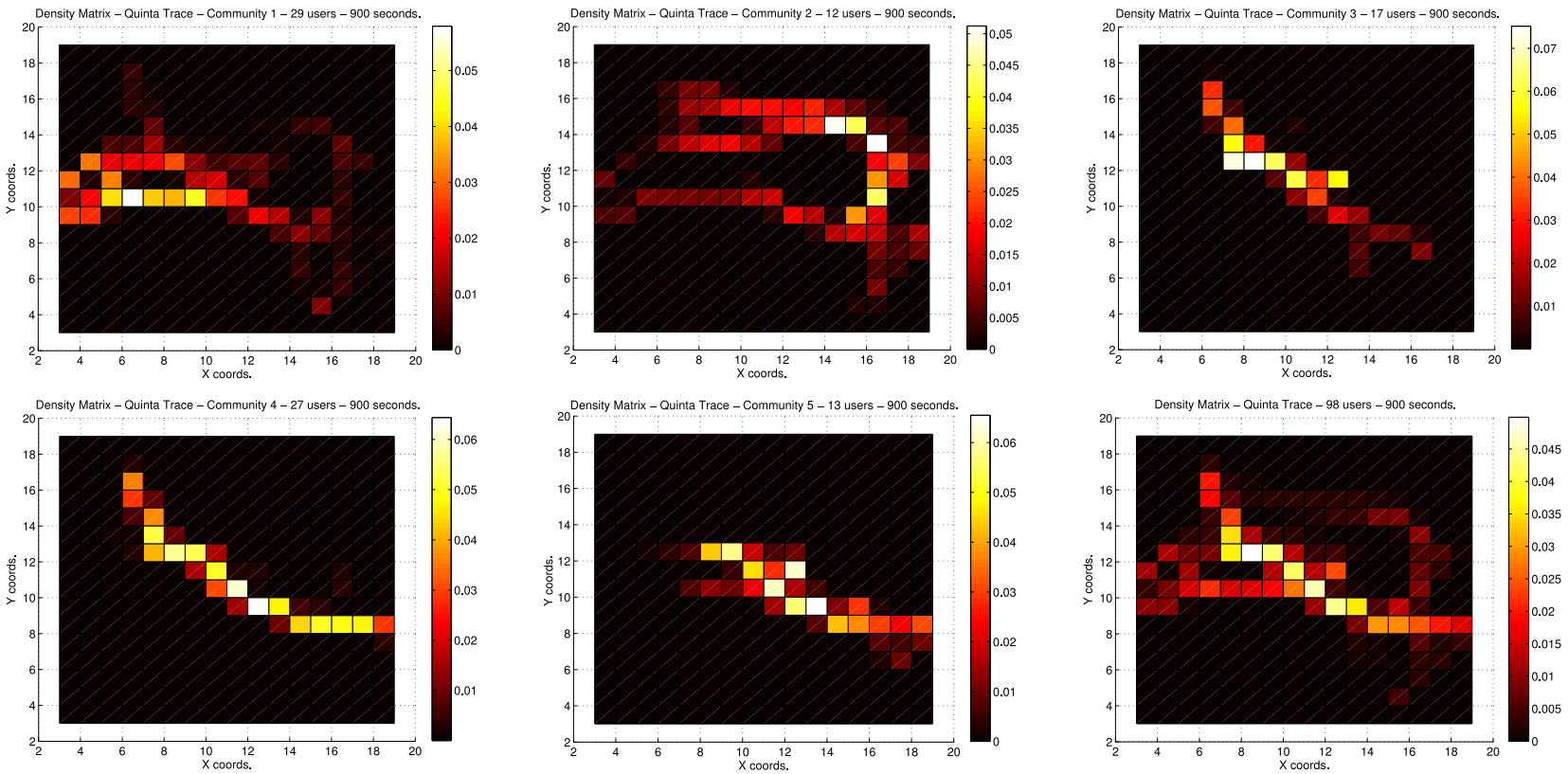
  - $\xi_n \in \{1,\ldots,K\}$ is the component indicator.
  - $a_{i,l,t-1}$ # of nodes from community l in node i at time $t-1$.
  - $\{c_{i,k}\}$ and $\{\theta_{k,l}\}$ are the unknown parameters to be estimated.
- The distributions of speeds and pause times are estimated using another DP mixture

Background

GPS trace collected at a park in the city of Rio de Janeiro, Brazil.

- The park has many trees, lakes, caves, and trails. It houses the National Museum of Natural History and the city Zoo.
- Total area covered is 840×840 meters.
- The analysis was performed on a 15 min (900 seconds) section of the full trace for which continuous traces (no missing values) is available for 98 nodes.
- Sampling frequency was 1 sec.
- Average speed is 1.27 m/s (± 0.16) and average pause time 4.61 s (± 1.60).

Clustering: Community-specific raw spatial intensities



Clustering: Community-specific speeds and pause times

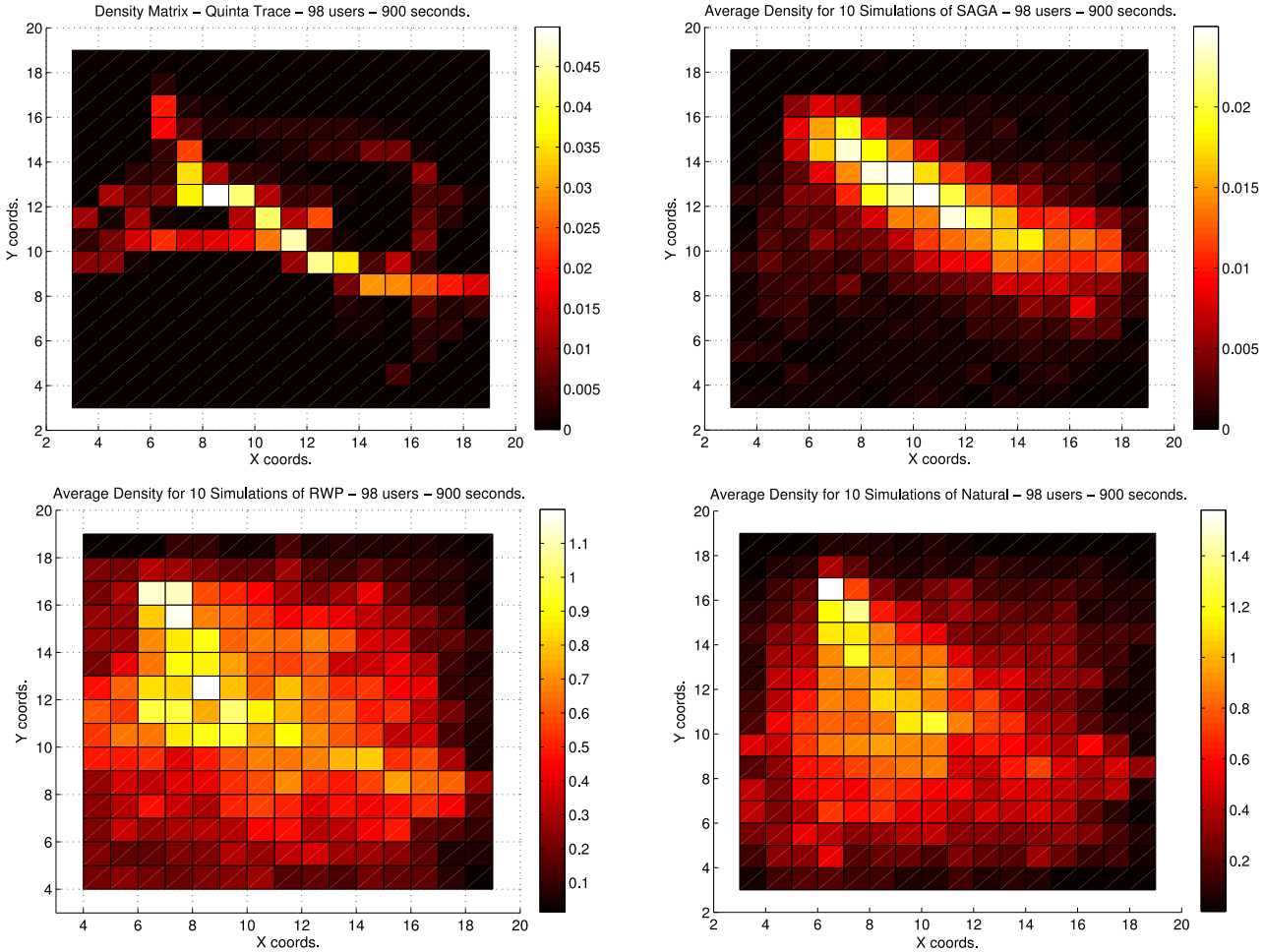
Community	Speed (m/s)	Pause Time (sec)
1	1.124 (1.117, 1.132)	3.316 (2.303, 4.330)
2	1.453 (1.444, 1.462)	2.365 (1.356, 3.375)
3	1.028 (1.017, 1.038)	5.951 (1.974, 9.928)
4	1.139 (1.132, 1.146)	2.036 (1.432, 2.640)
5	1.103 (1.092, 1.114)	4.320 (3.092, 5.548)

Multinomial regression: point and interval estimates

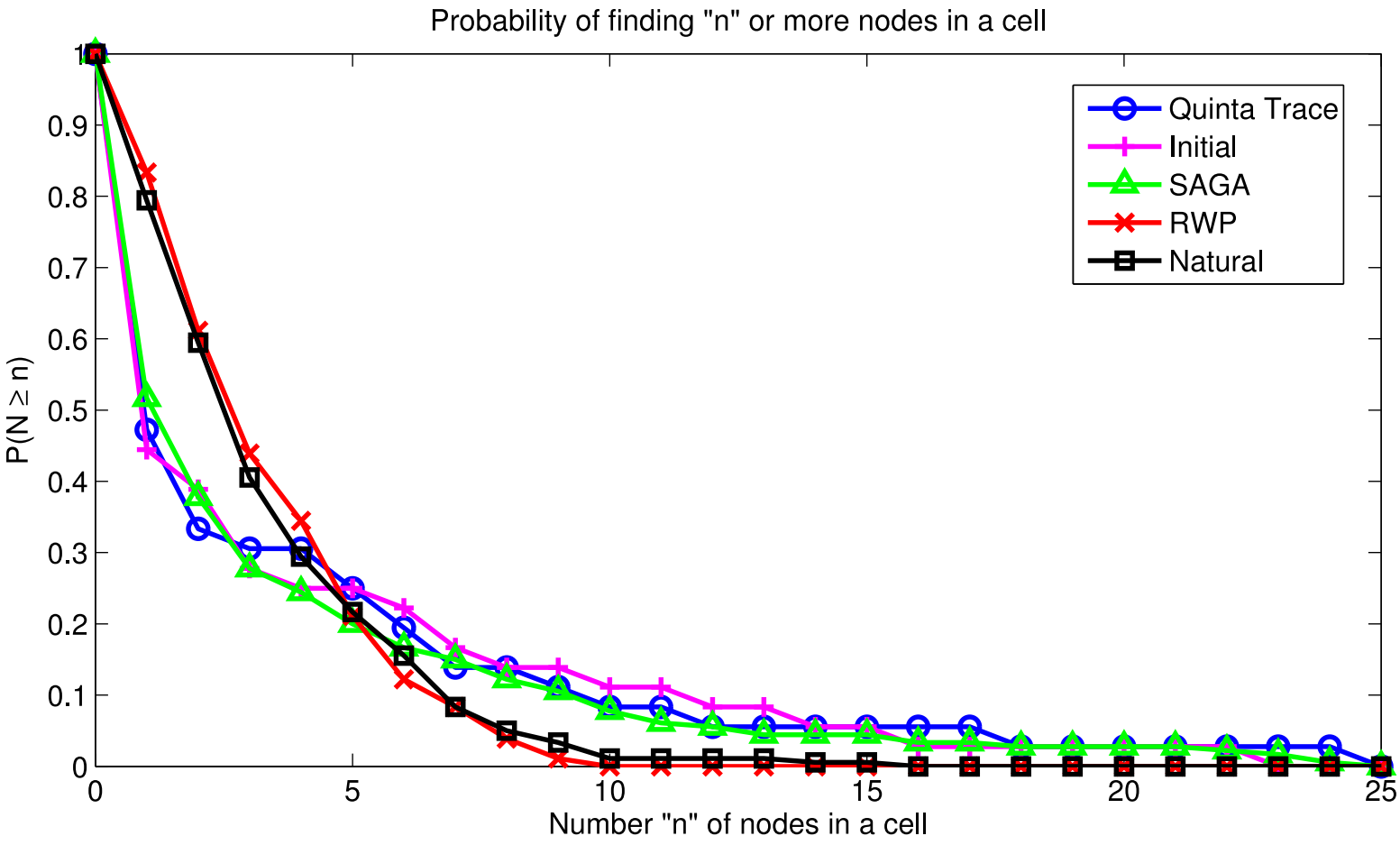
Θ	1	2	3	4	5
1	**-0.138** (-0.258,-0.012)	**0.299** (0.239,0.363)	**-0.135** (-0.21,-0.071)	**0.369** (0.307,0.495)	**-0.072** (-0.123,-0.008)
2		**-0.411** (-0.479,-0.341)	**-0.004** (-0.061,0.039)	**0.425** (0.357,0.497)	**0.151** (0.110,0.205)
3			**0.183** (0.113,0.269)	**0.494** (0.417,0.561)	**0.339** (0.284,0.374)
4				**0.512** (0.405,0.657)	**0.452** (0.393,0.518)
5					**-0.077** (-0.133,-0.022)

- Both assortative (3,4) and dissasortative (1,2,5) communities.
- Interaction between communities 2 and 3 appears not to be significant.

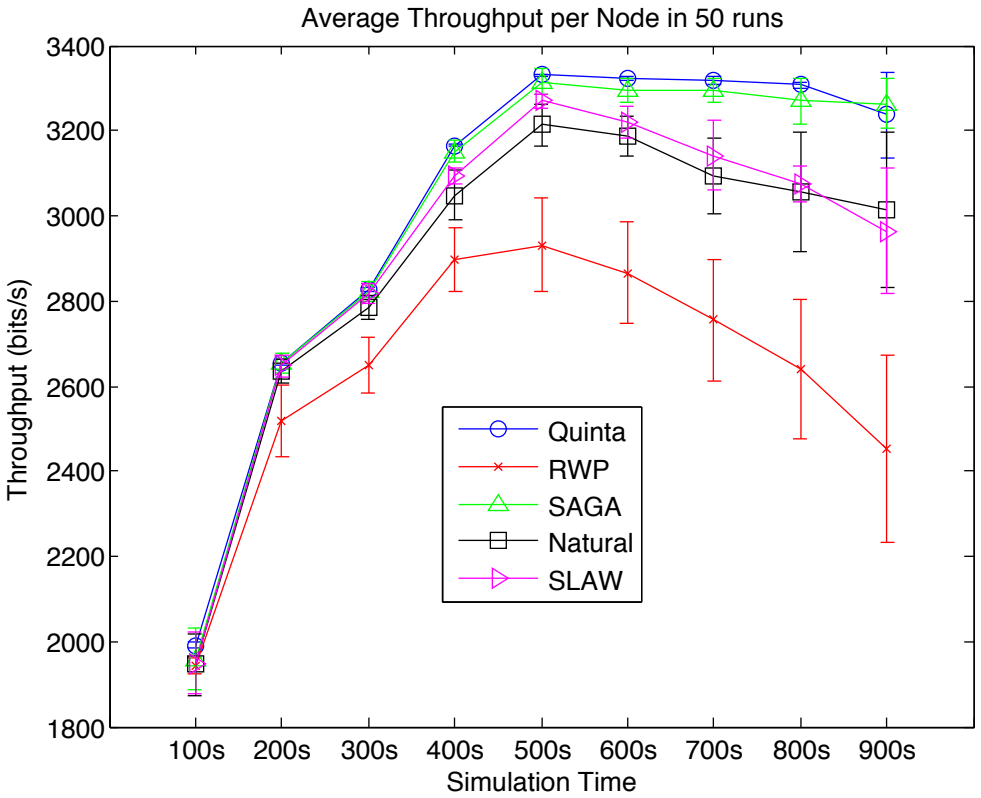
Forward simulation: Raw and simulated spatial intensities



Forward simulation: Node intensity

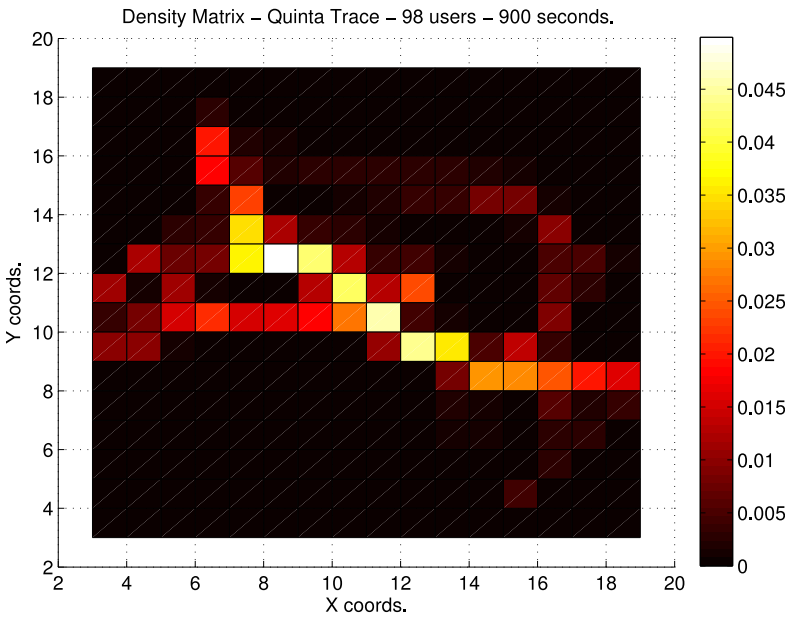


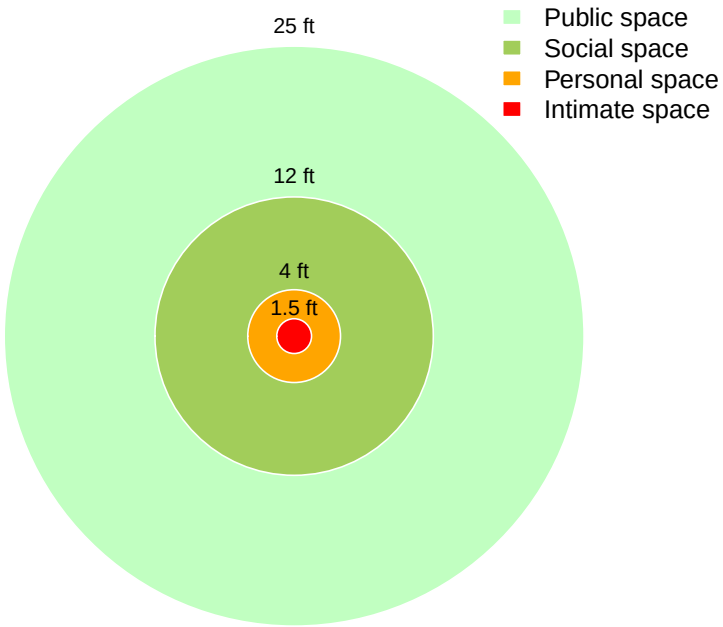
Forward simulation: Throughput



Ad-Hoc On-Demand Distance Vector Routing (AODV) [38] protocol, one of the Internet standards for MANET routing.

Some stylized features of human behavior

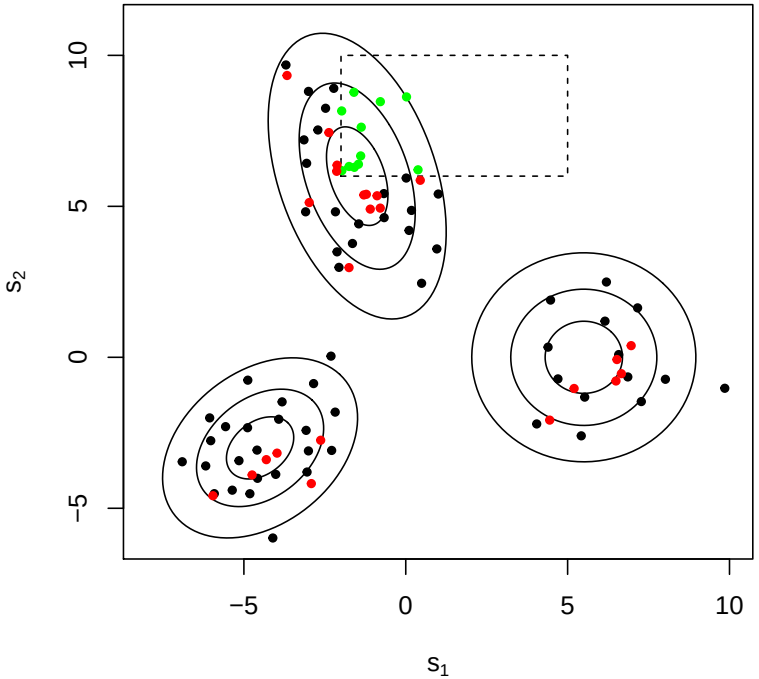




- Models that can capture stylized features?
- Dealing with challenges inherent to GPS data?

Matérn point processes

- Hardcore Matérn processes can capture clustering as well as repulsion.
- They are obtained by thinning a (non-homogeneous) Poisson process. Different "types" (I, II and III) according to how points are thinned.
- Type III: points in the primary process are assigned marks, points with later marks within a certain distance are removed.



A population-bsed model

- Model position of the individuals observed at time t with a Matérn process with primary intensity $\Lambda_t(\mathbf{s}) : \mathcal{S} \subset \mathbb{R}^2 \rightarrow \mathbb{R}^+$ and common thinning radius R for all individuals and times.
- Model the intensity using a nonparametric mixture prior

$$\Lambda_t(\mathbf{s}) = \sum_{k=1}^{\infty} w_{t,k} \Psi\left(\mathbf{s} \mid \boldsymbol{\theta}_k\right)$$

  A number of options are available but we are likely to focus on mixtures of Bernstein polynomials.
- Link intensities over time by introducing a stochastic model on the weights $(w_{t,1}, w_{t,2}, \ldots)$.
- This model tracks the behavior of the population, not the individuals.

An individual based model

- Consider the location of individual i at time t, $\tilde{\mathbf{s}}_{t,i}$. The evolution of the positions can be modeled using a mixture of autoregressive models,

$$f(\tilde{\mathbf{s}}_{t+1,i} \mid \tilde{\mathbf{s}}_{t,i}) = \sum_{k=1}^{\infty} \pi_k(\tilde{\mathbf{s}}_{t,i}) \, \mathrm{N}\left(\tilde{\mathbf{s}}_{t+1,i} \mid \boldsymbol{\mu}_k + \rho_k \left\{\tilde{\mathbf{s}}_{t,i} - \boldsymbol{\mu}_k\right\}, \boldsymbol{\Omega}_k\right).$$

  Individuals move towards attractor k with probability $\pi_k(\tilde{\mathbf{s}}_{t,i})$. The speed is controlled by $\boldsymbol{\Omega}_k$.
- Initial position comes from another mixture model.
- Gaps in the trace treated as missing values. Potential informative missingness.

Discussion

- Point process models can be helpful in understanding and predicting human mobility.
- Models need to be able to capture features of real data.
- Very interesting work ahead of us.