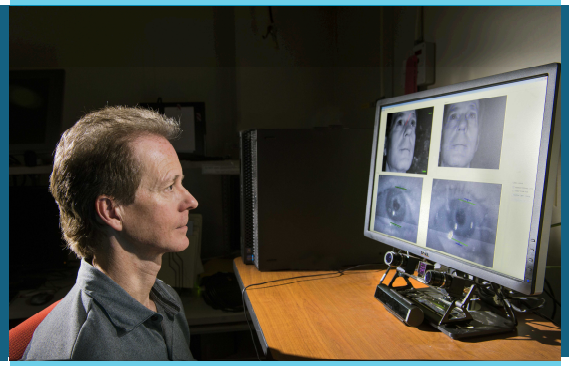
SAND2018-7699C

VISUALIZING CLUSTERING AND UNCERTAINTY ANALYSIS OF MULTIVARIATE TIME-SERIES DATA



PRESENTED BY

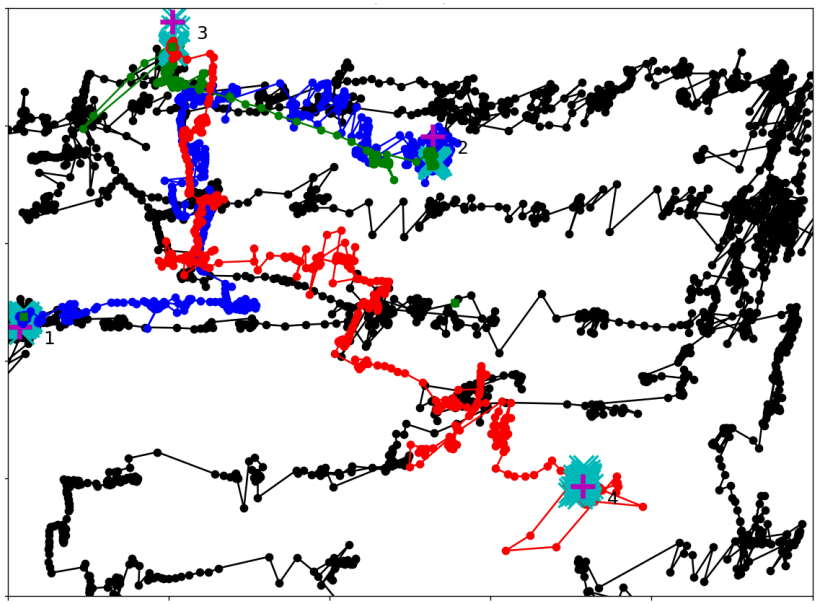
Kristin Divis, Maximillian Chen,

Laura McNamara, J. Dan Morrow

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Inspiration: Eye tracking data

- Series of gaze points (x, y) every ~17 ms—it's both **longitudinal** and **multivariate** (not i.i.d.)
- Questions:
  - How do we determine patterns of longitudinal, multivariate eye movement behavior in an **unsupervised** manner?
  - How do we quantify the **uncertainty** of this pattern determination in order for us to determine how confident we should be in the clustering results?
  - Can **visualizations** (in addition to global numerical measures) help build our understanding?



Probabilistic Clustering Models

- Provide probabilistic information about assignment of data points to clusters
  - Allow for **uncertainty quantification**

- Commonly-used probabilistic models such as the Gaussian Mixture Model (GMM)[1] and Latent Dirichlet Allocation (LDA)[2] assume data is **i.i.d.**

- The GMM has been extended for **scalar** longitudinal data[3] … but we are interested in models that also cluster **dependent multivariate** data

[1] Raftery et al. (2002)
[2] Blei et al. (2003)
[3] McNicholas et al. (2010)

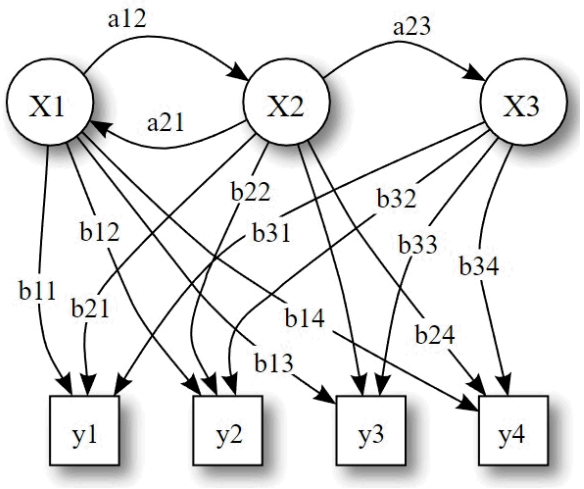
Hidden Markov Model (HMM)

Model Assumptions

- **Observed data**: *m*-variate time series of length *T* denoted by the general form: :

$$O_{1:T} = (O_1^1, \ldots, O_1^m, O_2^1, \ldots, O_2^m, \ldots, O_T^1, \ldots, O_T^m)$$

- **Latent (hidden) states**: $S_{1:T} = (S_1, \ldots, S_T)$
- **Model parameters**: θ
- **Covariates**: $z_{1:T} = (z_1, \ldots, z_T)$



X: hidden states
y: observed states
a: transition probabilities
b: emission probabilities

HMM: Uncertainties

Uncertainty Quantification: We want to quantify the uncertainty of the predicted state of an observation at time t

- **Posterior probability** of being in state j at time t given the observation sequence $O_{1:T}$, covariates $z_{1:T}$, and model parameters θ:

$$P(S_t = j \mid O_{1:T}, z_{1:T}, \theta')$$

- **State classification**: $S_t^* = \max_j P(S_t = j \mid O_{1:T}, z_{1:T}, \theta')$

- **Classification uncertainty**: $1 - \max_j P(S_t = j \mid O_{1:T}, z_{1:T}, \theta')$

Clustering Evaluation Measure: Numerical

- These are completely global measures: there is a **single numerical value** for the **entire data set**
  - All measures take values between 0 and 1 (0 = completely dissimilar; 1 = perfectly similar)

- **External Evaluation**: determine whether two clustering models produce similar clusters
  - Rand Index (RI)
  - Hubert and Arabie's Adjusted Rand Index
  - Morey and Agresti's Adjusted Rand Index
  - Fowlkes-Mallows (FM) Index
  - Jaccard Index (J)

- **Internal Evaluation**: identify separability of clusters
  - Dunn Index (D)

$$RI = \frac{TP + TN}{TP + FP + FN + TN},$$

where TP=true positive, TN=true negative, FP=false positive, and FN=false negative

$$ARI = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

$$FM = \sqrt{\frac{TP}{TP + FP}\frac{TP}{TP + FN}}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN},$$

where $|A \cap B|$ is the size of the intersection of datasets A and B and $|A \cup B|$ is the size of the union of datasets A and B.

$$D = \frac{\min_{1 \le i < j \le n} d(i, j)}{\max_{1 \le k \le n} d'(k)},$$

where d(i,j) represents the distance between clusters i and j, and d'(k) measures the intra-cluster distance of cluster k

Clustering Evaluation Measure: Visualizations

Expect visualizations to **enhance our understanding** of the clustering models

- Visualize clustering results and clustering uncertainty for a model
- Compare results of multiple models
- Provide **more specific information** on clustering trends than existing global numerical clustering evaluation measures

Data set

Constrained visual search task on synthetic aperture radar (SAR) imagery

- Created as a data "sandbox" to validate newly developed algorithms[1]
- Search for 4 target dots in a set order. Depending on task, may flip between different views of the same image to find the dots are make judgements between pairs of dots while searching.
- 16 participants, 4 task variants, took approximately 1 hour per participant
  - That's ~**25,000** sample points (x, y, t) per participant
  - … using a subset of the data here

Simplified view of task

[1] See Divis et al. (2018, NDIA)

Process

- Model fitting and selection
  - Use R package depmixS4 (Visser and Speenbrink, 2010)
  - Assume each (x,y) data point follows a multivariate normal distribution.
  - Select model using BIC criterion (lowest BIC value after fitting models with different numbers of hidden states)
- Covariates: featurize scanpath (e.g., changes in direction, curviness of path between targets)
- Model variants
  - **No** covariates
  - **All** covariates (length ratio, angle, angle difference, total angles)
  - **Single** best covariate (angle)
- Model evaluation
  - Global **numerical** external and internal measures
  - Cluster **visualizations**

Numerical Clustering Evaluation Results

External Evaluation: models with **no covariates** and **single angle covariate** are most similar (generally highest values)

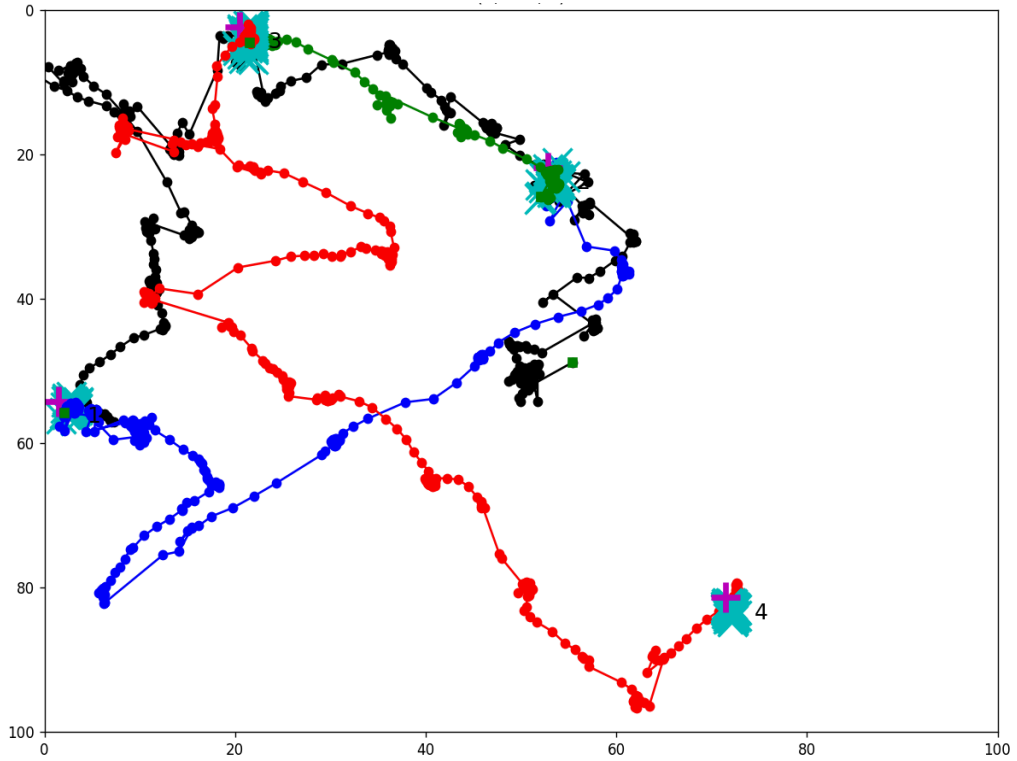
Model 1 Covariates	Model 2 Covariates	Rand	HA	MA	FM	Jaccard
None	**Angle**	0.859	**0.365**	**0.367**	**0.446**	**0.286**
None	Multiple	**0.865**	0.346	0.349	0.422	0.268
Angle	Multiple	0.840	0.280	0.283	0.372	0.228

Internal Evaluation: model with the **single angle covariate** has the highest separability (highest value)

Covariates	Dunn Index
None	0.00075
Angle	**0.00112**
Multiple	0.00012

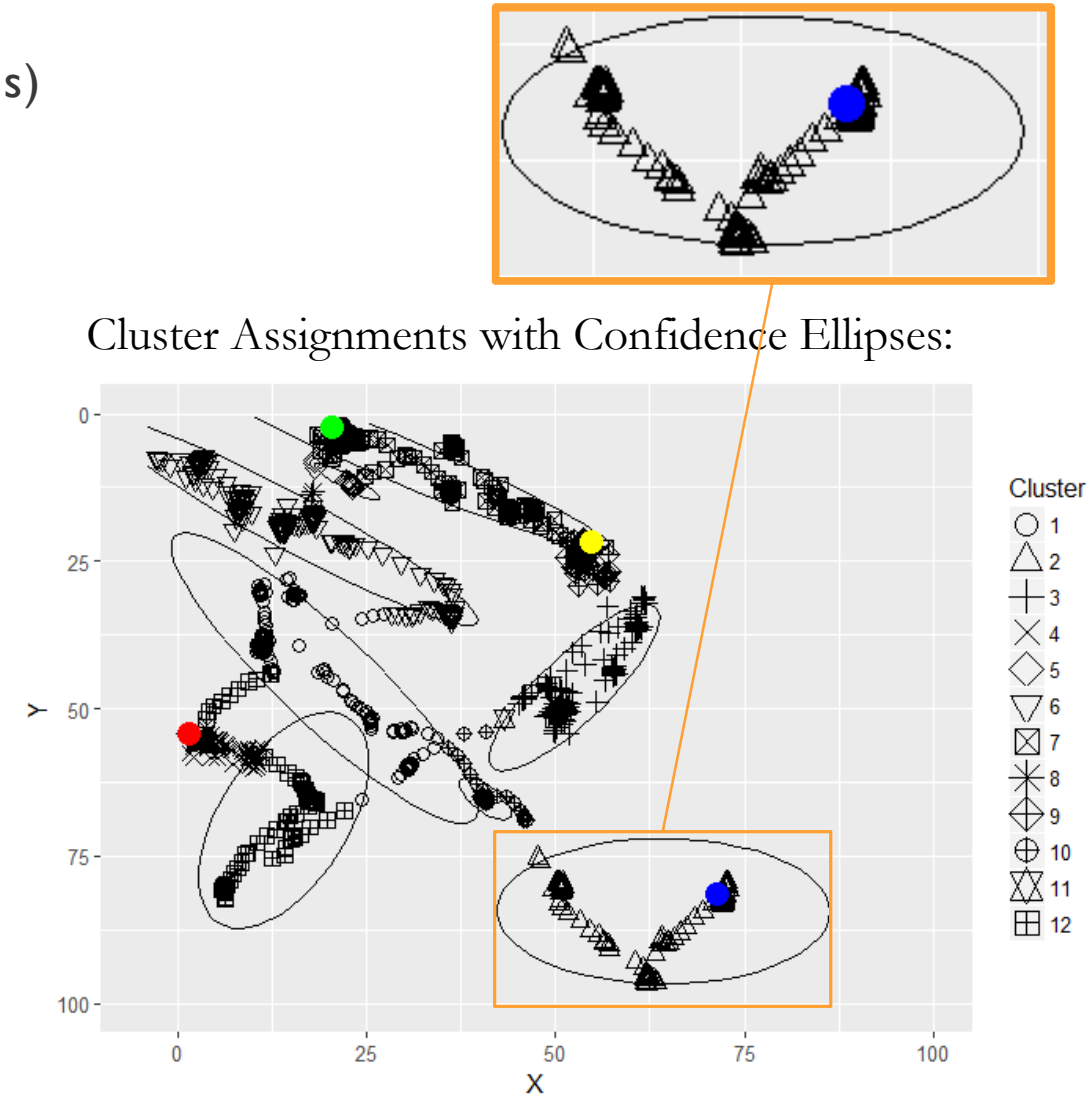
HMM with **no** covariates (12 Hidden States)

Scanpath:



Black: Start to Target 1 | Blue: Target 1 to Target 2
Green: Target 2 to Target 3 | Red: Target 3 to Target 4

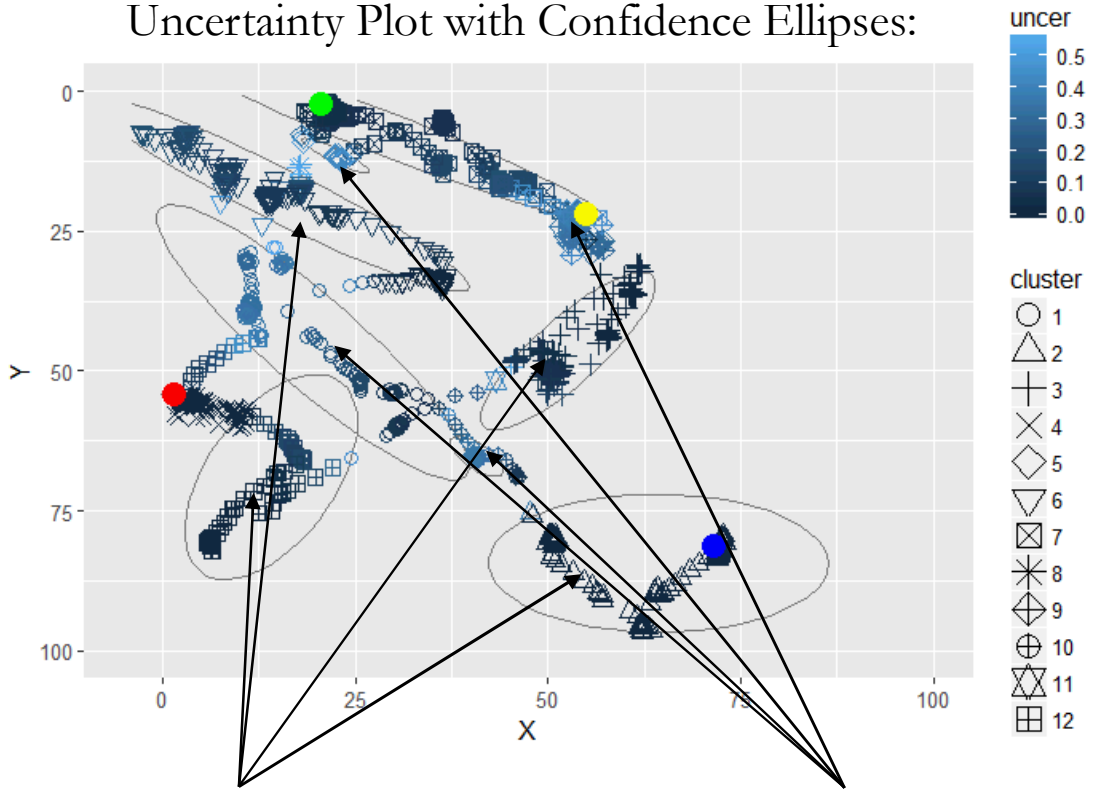
Cluster Assignments with Confidence Ellipses:



Relatively **distinct** and **tightly-packed** clusters

HMM with **no** covariates (12 hidden states)

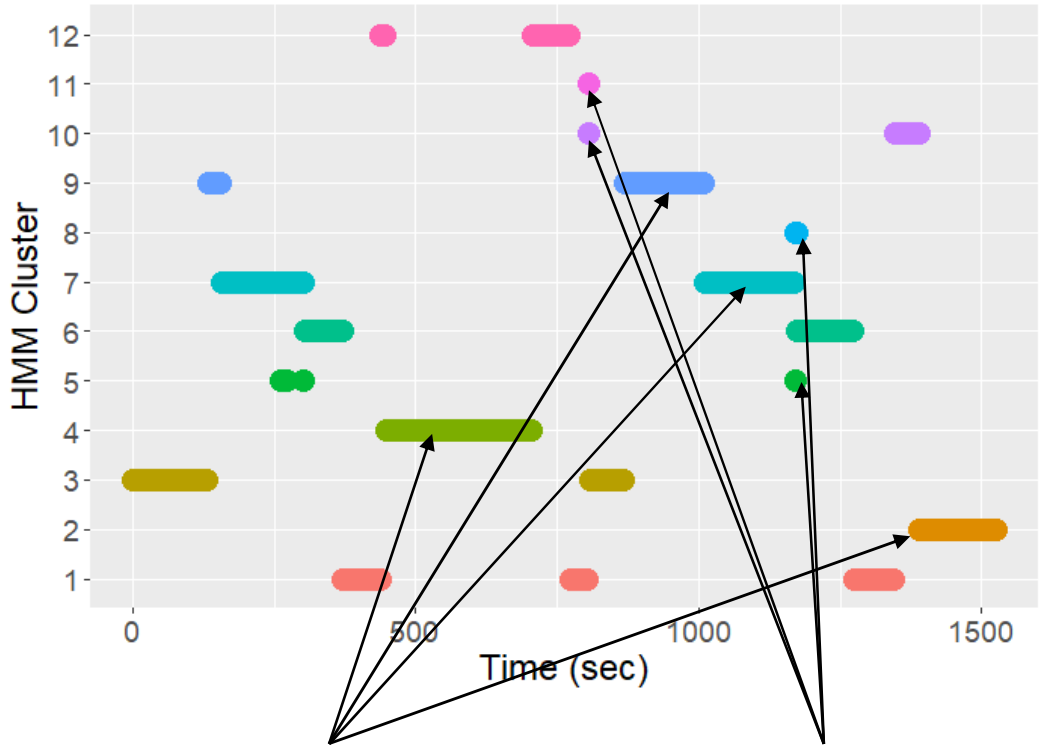
Uncertainty Plot with Confidence Ellipses:



Clusters 2, 3, 6, & 12 all have points classified to these clusters with **low uncertainty**

Clusters 1, 5, 9, & 10 all have points classified with relatively **high uncertainty**

Time Plot:

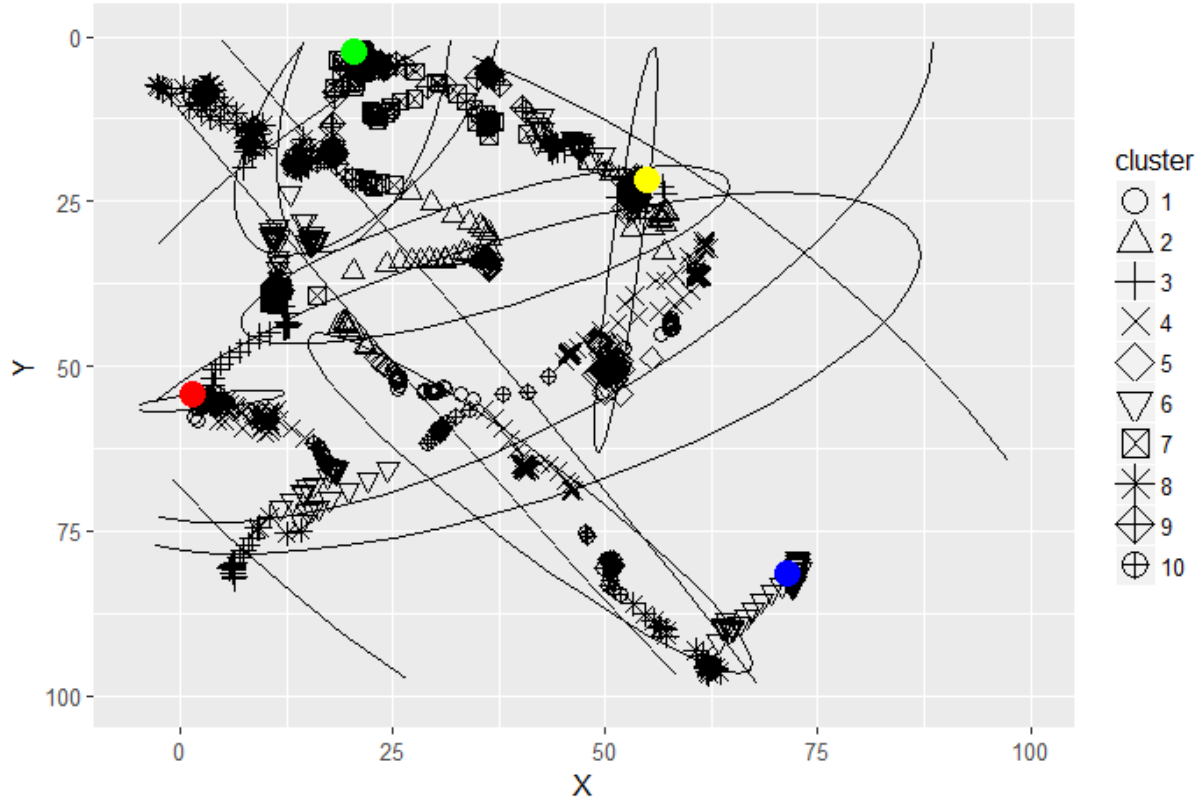


Participants spend relatively **long periods of time** in some clusters

…but relatively **little time** in other clusters

HMM with **all** covariates (10 hidden states)

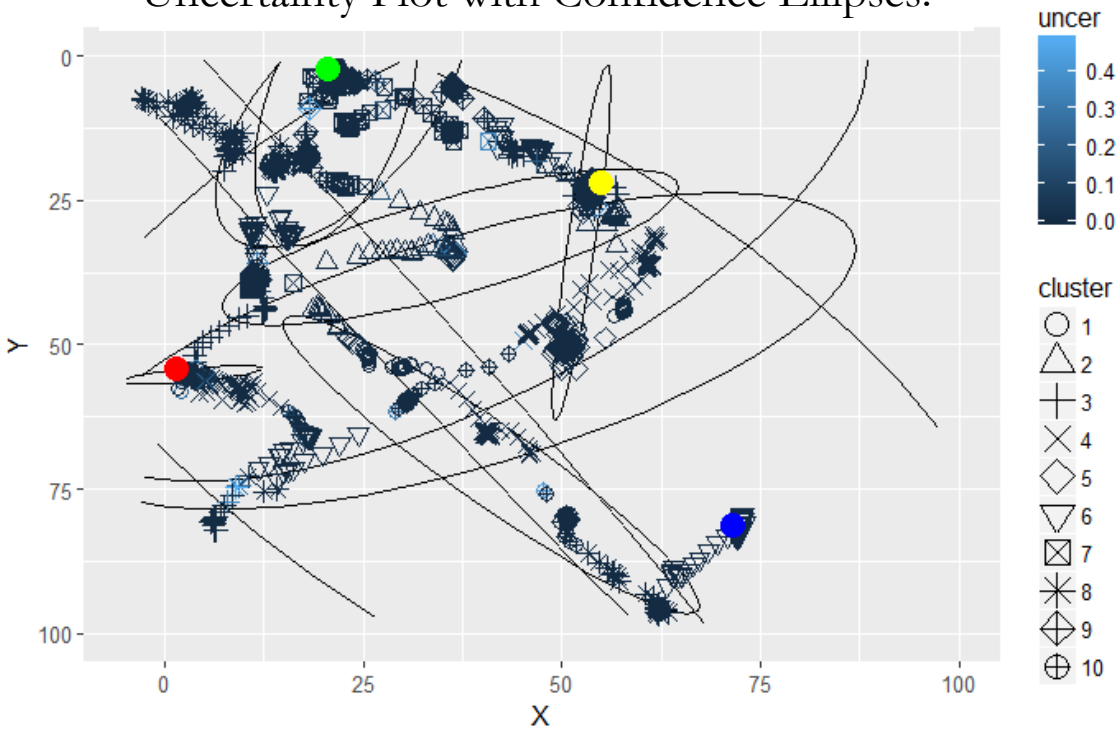
Cluster Assignments with Confidence Ellipses:



Adding all the covariates leads to **wider** clusters and **more overlap** between clusters—does not appear to improve clustering results

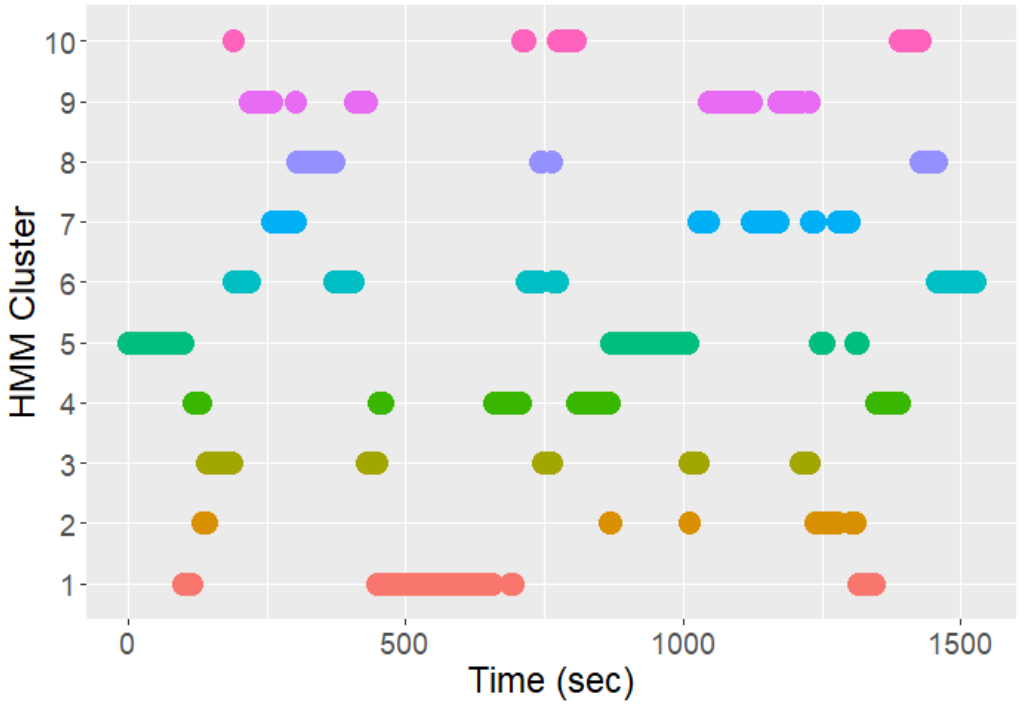
HMM with **all** covariates (10 hidden states)

Uncertainty Plot with Confidence Ellipses:



Most of the points have been assigned to clusters with **relatively low uncertainty** but there's still **large overlap** between cluster ellipses, leading to overall poor confidence in the clusters

Time Plot:

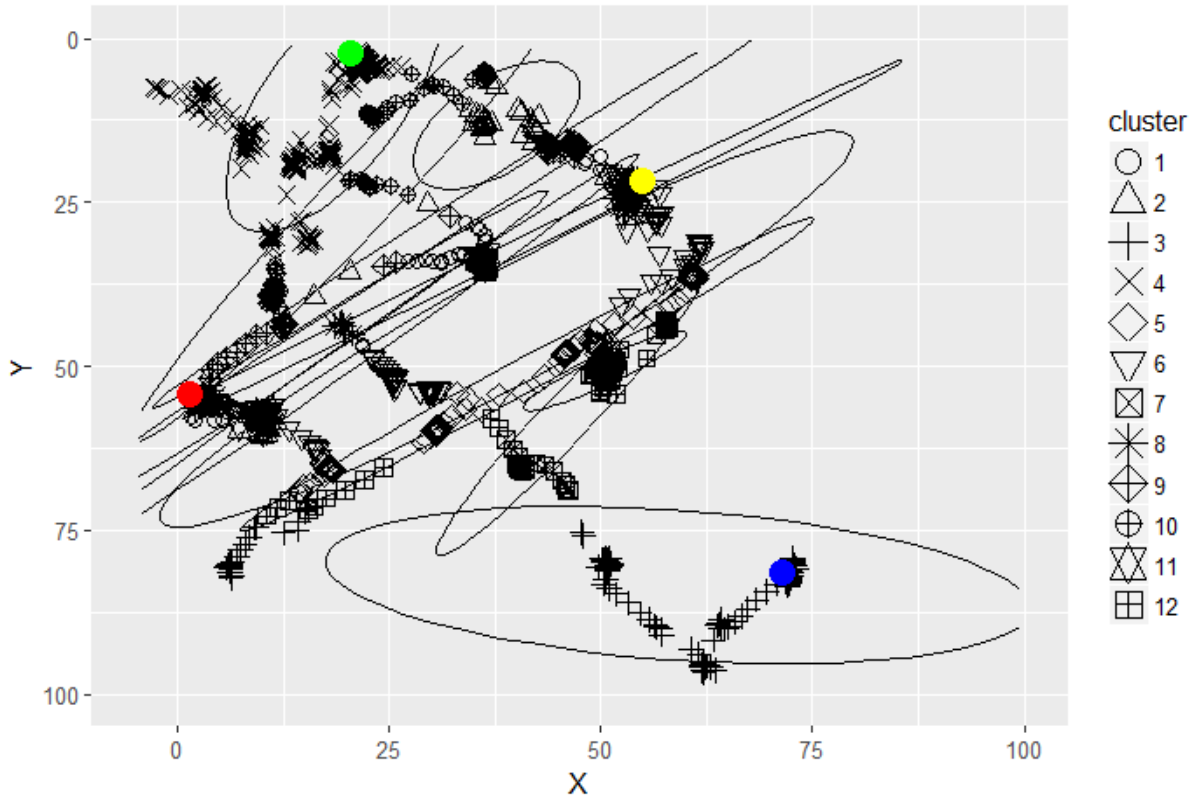


Similar pattern of some clusters having relatively **long dwell times**

…but also have more instances of **short revisits** to the same cluster

HMM with **angle** covariate (12 hidden states)

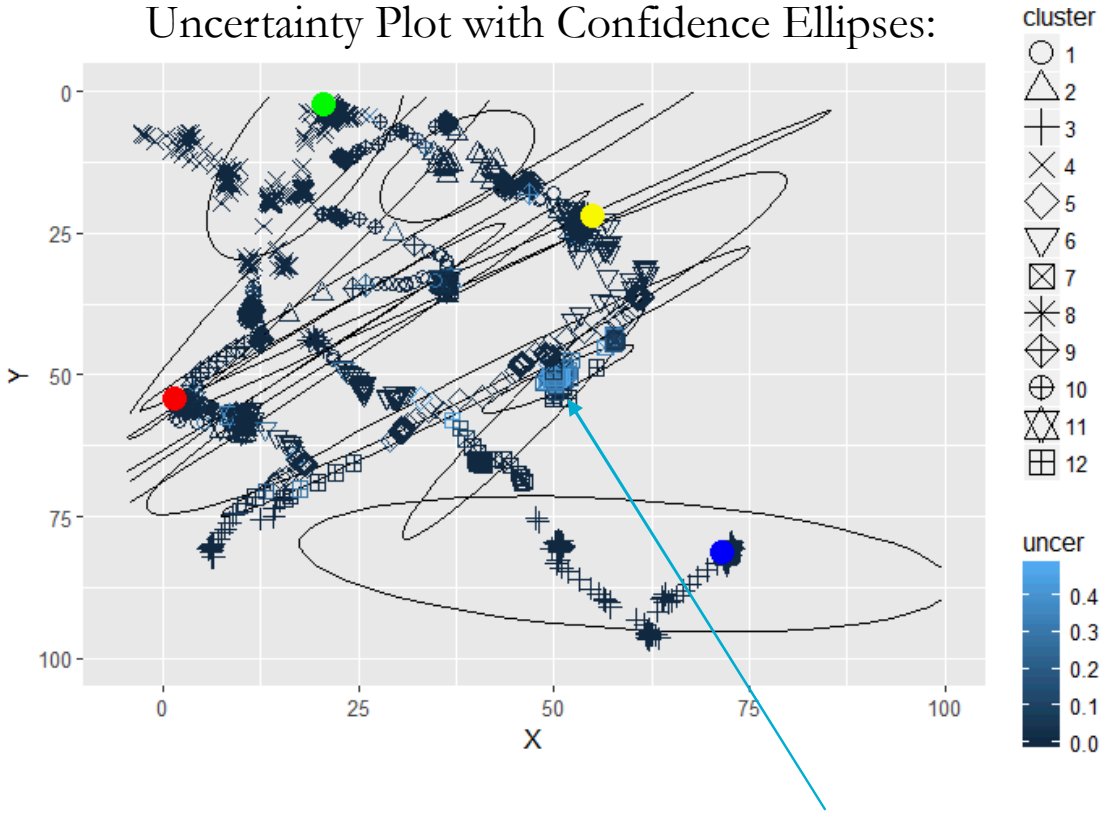
Cluster Assignments with Confidence Ellipses:



Having the angle covariate alone is better than all covariates (fewer overlapping clusters) but the clusters are still not as distinct as when no covariates are included

HMM with **angle** covariate (12 hidden states)

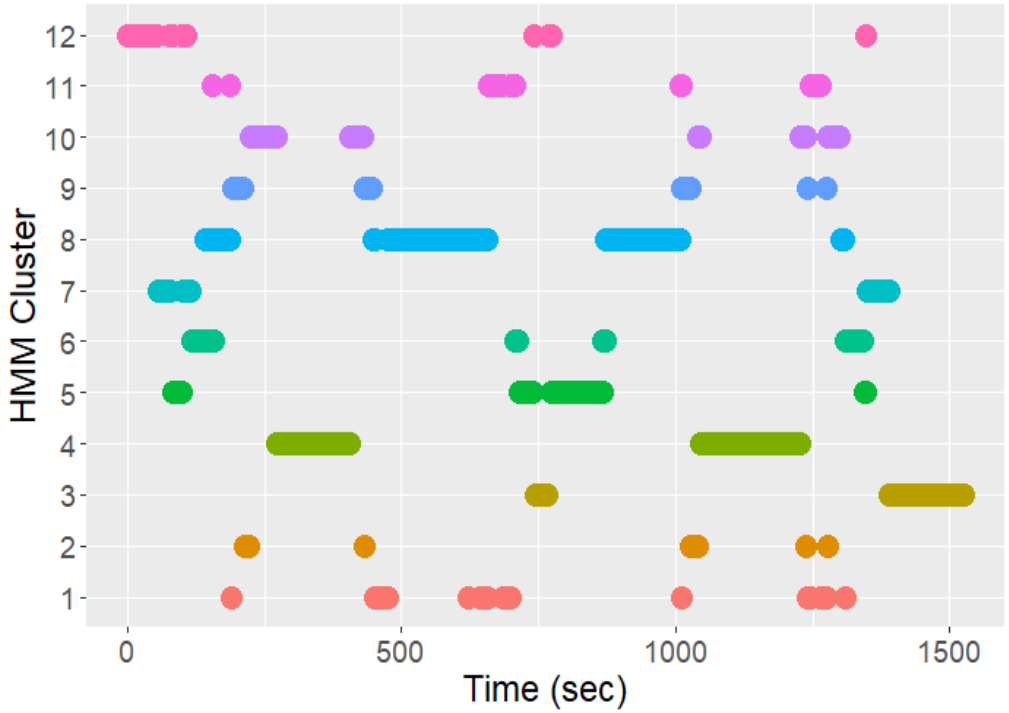
Uncertainty Plot with Confidence Ellipses:



Most points clustered with relatively **low uncertainty** (dark blue)

…the **high uncertainty** (light blue) points are mainly around (50,50) in cluster 12

Time Plot:



Spend a relatively large amount of time in some clusters and very little time in others, with multiple short revisits

Summary

- Used **HMM** methods to cluster **multivariate**, **time-dependent** data (multiple models with levels of covariate inclusion) and evaluated with global **single-value numerical** clustering measures and **visualizations**
- Inspired by eye tracking data—but models did **not** fit that data well
  - Including visualizations as an evaluation tool allowed us to better understand the **why** and **how**
    - Cluster assignments with 95% confidence ellipse: allowed us to gauge the size an variability of each cluster
    - Cluster assignments with uncertainty of each data point coded via color: allowed us to gauge uncertainty relative to position in cluster
    - Time plot of clusters: allowed us to track the assignment of data points through time
- Benefits of visualization:
  - Assess the clustering performance and what level of confidence we should have in the results
  - Identify specific patterns in the data, which existing numerical clustering evaluation measures cannot provide
  - Compare results of multiple clustering models

Future Work

- **Statistical**
  - Extend single-numerical evaluation measure and visualizations to data sets and models with better clusters
  - Goodness-of-fit statistics for mixture models
  - Integrate cluster separability measures into the computation of classification uncertainty
  - Extend current measures for classification uncertainty at individual data points to quantifying the uncertainty of clusterings, and then visualizing these uncertainty bounds
  - Create visualizations for clustering time-series data with data points of more than two dimensions
- **Geospatial temporal** (eye tracking) data: more *useful* approach?
  - Currently every data point is forced into a cluster—traditional eye tracking techniques **drop data** that do not align to meaningful eye movement patterns
  - Pull in **top-down components** (e.g., fixations, saccades, blinks) to help guide the bottom-up clustering into more meaningful patterns?
  - Better incorporate **temporal** information in addition to spatial (current model only "looks" 1 point back)
  - Apply to less coarse eye movement data (e.g., letter shapes in reading)

Thank you!

Questions?

Kristin Divis (kmdivis@sandia.gov) | Max Chen (mgchen@sandia.gov) | Laura McNamara (lamcnam@sandia.gov) | Dan Morrow (jdmorr@sandia.gov)