SANDIA REPORT

SAND2018-10359
Unlimited Release
Printed September, 2018

LDRD PROJECT NUMBER: 209227 LDRD PROJECT TITLE: Visualizing Clustering and Uncertainty Analysis with Multivariate Longitudinal Data

Maximillian G. Chen, Kristin M. Divis, J. Dan Morrow, and Laura A. McNamara

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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 Energys National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2018-10359
Unlimited Release
Printed September, 2018

LDRD PROJECT NUMBER: 209227 LDRD PROJECT TITLE: Visualizing Clustering and Uncertainty Analysis with Multivariate Longitudinal Data

Maximillian G. Chen
Data Science and Applications Department
Sandia National Laboratories
mgchen@sandia.gov

Kristin M. Divis
Advanced Exploitation/Human-Systems Integration Department
Sandia National Laboratories
kmdivis@sandia.gov

J. Dan Morrow
Advanced Exploitation/Human-Systems Integration Department
Sandia National Laboratories
jdmorr@sandia.gov

Laura A. McNamara
Advanced Exploitation/Human-Systems Integration Department
Sandia National Laboratories
lamcnam@sandia.gov

1 ABSTRACT

Multivariate time-series datasets are intrinsic to the study of dynamic, naturalistic behavior, such as in the applications of finance and motion video analysis. Statistical models provide the ability to identify event patterns in these data under conditions of uncertainty, but researchers must be able to evaluate how well a model uses available information in a dataset for clustering decisions and for uncertainty information. The Hidden Markov Model (HMM) is an established method for clustering time-series data, where the hidden states of the HMM are the clusters. We develop novel methods for quantifying the uncertainty of the performance of and for visualizing the clustering performance and uncertainty of fitting a HMM to multivariate time-series data. We explain the usefulness of uncertainty quantification and visualization with evaluating the performance of clustering models, as well as how information exploitation of time-series datasets can be enhanced. We implement our methods to cluster patterns of scanpaths from raw eye tracking data.

2 INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called a *cluster*, consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering methods perform unsupervised learning of *hidden patterns* in the data. They have been used in many disciplines, such as statistics, pattern recognition, and image segmentation and computer vision. Clustering has been brought to life in data mining due to intense developments in information retrieval and text mining, spatial database applications (for example, GIS or astronomical data), sequence and heterogeneous data analysis, Web applications, and DNA analysis (6). There are also many classes of clustering algorithms. Hierarchical algorithms such as CURE (Clustering Using REpresentatives) learn clusters gradually (20), while partitioning methods, such as K-means (28), probabilistic clustering methods, and density-based methods such as OPTICS (1) and DBSCAN (11), learn clusters directly. There are also grid-based clustering methods (which perform space segmentation and then perform appropriate segment aggregation) such as GRIDCLUST (39), constrained-based clustering methods (that factor in problem-specific limitations) such as the COD (Clustering with Obstructed Distance (41)), and artificial neural network clustering such as SOM (Self-Organized Map) (25). See (6) for an overview of clustering techniques. While these methods have provided quality clustering results, with the exception of *probabilistic clustering* methods, they do not provide *probabilistic information*, i.e. the probabilities that a data point belongs in a specific cluster. Without this probabilistic information, we are unable to quantify the *uncertainty* of the clustering results and examine the *trustworthiness* of the results. This can be very important in areas such as national security (detecting threats) or medicine (detecting life-threatening diseases), where decisions can have far-reaching consequences.

Probabilistic clustering methods typically assume data to be sampled independently from a *mixture model* of several probability distributions. The main assumption is data points are generated by, first, randomly picking a model j with probability τ_j, $j = 1, ..., K$, where K is the number of models, and second, by drawing a point x from a corresponding distribution. The area around the mean of each (supposedly unimodal) distribution constitutes a natural cluster. Therefore, we associate each cluster with the corresponding distribution and its parameters. Notable probabilistic clustering methods include the Gaussian mixture model (GMM), where each cluster is assumed to follow a multivariate normal distribution with its unique mean vector and covariance matrix (15), and the non-parametric mixture model (NMM), where each cluster's distribution is not a well-defined parametric probability distribution. Rather, each cluster's density is estimated via kernel density estimation (5; 4). All of these methods assume that each data point is independent and identically distributed (i.i.d.), which means they are not directly applicable for datasets containing dependent observations. Visualizing the clustering and uncertainty results for fitting a GMM to i.i.d. data is provided in the MCLUST algorithm and its associated R package (15; 16; 40). The GMM has been extended for univariate longitudinal data, where a data vector consists of scalar observations that are temporally correlated, but the data vectors are i.i.d (33). Thus, their methods do not apply to datasets consisting of dependent *vectors*, where each n-dimensional vector must be considered as an entire observation and is not decomposed into its scalar components. Longclust, the associated method and R package of (33), does not provide visualizations for the clustering and uncertainty results.

Classification uncertainty, which looks at the probability a data point is not classified properly, has been investigated for the GMM with i.i.d. data. For a GMM, the misclassification probability between two clusters i and j has been computed by (30; 34) as

$$\omega_{j|i} = \Pr(\tau_i \phi(\mathbf{Y}; \mu_i, \mathbf{\Sigma}_i) < \tau_j \phi(\mathbf{Y}; \mu_j, \mathbf{\Sigma}_j) | \mathbf{Y} \sim N_p(\mu_i, \mathbf{\Sigma}_i)), \tag{1}$$

where $\phi(\cdot)$ is the probability density function of the multivariate normal distribution. The general classification uncertainty of observation i has been computed by (15) as

$$1 - \max_k z^*_{ik}, \tag{2}$$

where z^*_{ik} is the estimate of the posterior probability of observation i belonging in cluster k, as estimated by the EM algorithm.

The Hidden Markov Model (HMM) is an established method for clustering time-dependent data (18). They have been prominently used in speech recognition (12), but they have also been applied to other applications, such as in the social sciences (35; 36; 46; 27), biology (26), econometrics (24), and machine learning and data mining (17). See

(42) for a more detailed summary of the use of HMMs in these areas. These models have been used in psychological experiments for modeling state-switching processes occurring within a person, dyad, family, group, or other system over time (7). The HMM has also been used for clustering spatio-temporal data in application such as detecting brain changes in brain imaging data (45), conflict data (47), land use (32; 19), and genomic data (31). In all of these applications, the clustering is investigated in depth and the temporal changes in the clusterings are visualized. However, none of these works address the classification uncertainty of the HMM to their respective datasets.

The paper is organized as follows. In section 3, we will describe the formulation of the HMM, as well as our method for quantifying the uncertainty of the clustering performance of the HMM and visualizing this performance. In section 4, we describe the application of HMMs to clustering eye tracking data, which contain spatio-temporal information on where a person's eye looks at and at what time during a visual search task. We will conclude the paper with discussion and conclusions in section **??**.

3 DETAILED DESCRIPTION OF EXPERIMENT/METHOD

3.1 Hidden Markov Model

The Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. Unlike simpler Markov models, such as a Markov chain, where the state at a given time point t is observed, in a HMM, the state is unobserved, or *hidden*. Instead, the output, which is dependent on the hidden state, is observed.

Formally, a HMM has the following components:

1. **Observed Data**: an m-variate time series of length T denoted by the general form

$$\mathbf{O} = \mathbf{O}_{1:T} = (O_1^1, ..., O_1^m, O_2^1, ..., O_2^m, ..., O_T^1, ..., O_T^m) \tag{3}$$

2. **Latent (hidden) states**: $\mathbf{S}_{1:T} = (S_1, ..., S_T)$, which belong to a finite state space $\mathcal{S} = \{1, ..., n\}$, a set of n latent states.

3. **Model parameters**: θ

4. **Covariates**: $\mathbf{z}_{1:T} = (\mathbf{z}_1, ..., \mathbf{z}_T)$

5. **Transition Probabilities** $a_{ij}(\mathbf{z}_t) = P(S_{t+1} = j|S_t = i, \mathbf{z}_t)$: the probability of a transition from state i to state j with covariate $\mathbf{z}_t$.

6. **Observation likelihoods** or **emission probabilities** b_{S_t}: a vector of observation densities $b_j^k(\mathbf{z}_t) = P(O_t^k|S_t = j, \mathbf{z}_t)$ that provide the conditional densities of observations O_t^k associated with latent class/state j and covariate $\mathbf{z}_t$, $j = 1, ..., n$, $k = 1, ..., m$.

7. Initial state probabilities $\pi_i(\mathbf{z}_i) = P(S_1 = i|\mathbf{z}_1)$: the probability of class/state i at time $t = 1$ with covariate $\mathbf{z}_1$.

The HMM contains two assumptions (23; 42). First, the hidden states follow the Markov property, i.e. the hidden state at time t, S_t, is only dependent on the hidden state at time $t-1$, S_{t-1}. Formally, this is written as

$$P(S_t|S_{t-1}, S_{t-2}, ..., S_1) = P(S_t|S_{t-1}). \tag{4}$$

Let s_i denote the actual hidden state at time i. The second assumption is the probability of an observation o_i depends only on the hidden state s_i and not on any other hidden states or observations:

$$P(o_i|s_1, ..., s_i, ..., s_T, o_1, ..., o_i, ..., o_T) = P(o_i|s_i). \tag{5}$$

3.2 Likelihood Functions and Parameter Estimation

The joint likelihood of observations $\mathbf{O}_{1:T}$ and latent states $\mathbf{S}_{1:T} = (S_1, ..., S_T)$, given model parameters θ and covariates $\mathbf{z}_{1:T} = (\mathbf{z}_1, ..., \mathbf{z}_T)$, can be written as

$$P(\mathbf{O}_{1:T}, \mathbf{S}_{1:T}|\theta) = \pi_i(\mathbf{z}_1)\mathbf{b}_{S_1}(\mathbf{O}_1|\mathbf{z}_1) \prod_{t=1}^{T-1} a_{ij}(\mathbf{z}_t) b_{S_t}(\mathbf{O}_{t+1}|\mathbf{z}_{t+1}). \tag{6}$$

To obtain the maximum likelihood estimates of the parameters, we need the marginal likelihood of the observations. For HMMs, this marginal (log-) likelihood can be computed by the forward-backward algorithm (2; 12). The forward algorithm is modified by (29) to allow computing of the gradients of the log-likelihoods for each observation at the same time. They rewrite the likelihood as

$$L_T = P(\mathbf{O}_{1:T}) = \prod_{i=1}^{T} P(\mathbf{O}_t|\mathbf{O}_{1:(t-1)}), \tag{7}$$

where $P(\mathbf{O}_1|\mathbf{O}_0) := P(\mathbf{O}_1)$. Note that for an observed Markov chain, these probabilities reduce to $P(\mathbf{O}_t|\mathbf{O}_1, ..., \mathbf{O}_{t-1}) = P(\mathbf{O}_t|\mathbf{O}_{t-1})$. The log-likelihood can now be expressed as

$$l_T = \sum_{t=1}^{T} \log[P(\mathbf{O}_t|\mathbf{O}_{1:t-1})]. \tag{8}$$

To compute the log-likelihood, (29) define the following forward recursion:

$$\phi_1(j) := P(\mathbf{O}_1, S_1 = j) = \pi_j \mathbf{b}_j(\mathbf{O}_1) \tag{9}$$

$$\phi_t(j) := P(\mathbf{O}_t, S_1 = j|\mathbf{O}_{1:(t-1)}) = \sum_{i=1}^{n} [\phi_{t-1}(i) a_{ij} \mathbf{b}_j \mathbf{O}_t] \times (\Phi_{t-1})^{-1}, \tag{10}$$

where $\Phi_t = \sum_{i=1}^{n} \phi_t(i)$. Combining $\Phi_t = P(\mathbf{O}_t|\mathbf{O}_{1:(t-1)})$, and equation (8) gives the following expression for the log-likelihood:

$$l_T = \sum_{t=1}^{T} \log \Phi_t. \tag{11}$$

The standard algorithm for fitting a HMM is the forward-backward, or Baum-Welch, algorithm (3), a special case of the Expectation-Maximization (EM) algorithm (8; 23; 42). This is an iterative algorithm that trains both the transition probabilities A and the emission probabilities B of the HMM. The R package depmixS4 uses the EM algorithm or the Newton-Raphson optimizer to estimate the parameters of the prior model, transition model, and response models (43).

3.3 Computing Posterior Probabilities and Uncertainty

Following the methods of (15), we write the **posterior probability** of the data point at time t being in state j given the observation sequence $\mathbf{O}_{1:T}$, covariates $\mathbf{z}_{1:T}$, and model parameters θ as

$$P(S_t = j|\mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta'). \tag{12}$$

The posterior probability is estimated via the Baum-Welch algorithm. After the posterior probability is computed, we can classify the data point and estimate the classification uncertainty of the data point.

The **state classification** at time t can be written as

$$S_t^* = \max_j P(S_t = j|\mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta'). \tag{13}$$

The **classification uncertainty** at time t is

$$1 - \max_j P(S_t = j|\mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta'). \tag{14}$$

3.4 Visualizations

We produce the following visualizations that allow us to see the clustering results and the clustering uncertainties of spatial location (the (x,y) coordinates) data indexed by time.

- **Clustering Results Plot**: We plot the spatial location of each data point with a *corresponding symbol to indicate the cluster the data point has been assigned to*.
- **Clustering Results Plot with Ellipses**: In addition to plotting the data points with the symbol corresponding to the point's assigned cluster, we also draw the 95% confidence ellipses for each cluster. This allows us to visualize the variance, and thus the range, of points assigned to each cluster. We can see whether or not a cluster is tightly bounded and is describing trends and behaviors within a constrained area.
- **Uncertainty Results Plot**: We plot the spatial location of each data point with both the corresponding symbol for the cluster the point has been assigned to and a *corresponding color for the classification uncertainty level*. Not only can we visualize which cluster a data point has been assigned to, but also what level of certainty that assignment has been made with.
- **Uncertainty Results Plot with Ellipses**: On top of the Uncertainty Results Plot, we draw the 95% confidence ellipses for each cluster. This plot allows us to visualize how the uncertainty results can change with the location of a data point within its assigned cluster. We can visually determine if data points at or near the center of the cluster have much *lower uncertainty*, meaning that the cluster assignment is made with much higher confidence, and if data points at the outer edges of the cluster have much *higher uncertainty*, which indicate the cluster assignment is made with much lower confidence.
- **Time Plot**: Since each data point is assigned to a cluster, the time plot allows us to track the cluster assignments in chronological order of the data points. We can see if clusters are "revisited," or if data points are assigned to a cluster from earlier in the data set, and if that's the case, how often clusters are "revisited."

3.5 Clustering Evaluation Measures

As a point of comparison for the visualizations, we will compute various existing clustering evaluation measures. These measures can be broadly divided into two categories. The first category is internal evaluation, when a clustering result is evaluated based on the data that is clustered itself. The second category is external evaluation, where clustering results are evaluated based on data that was not used for clustering. Below, we will describe the measures we compute.

For all of the measures, there is one quantitative result for the entire dataset. Therefore, these measures will not be able to distinguish patterns within the data that can be revealed by visualizing the clustering results. These measures are computed to supplement the analysis from the visualizations above.

3.5.1 Internal Evaluation

We compute the Dunn index, which aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index is calculated as

$$D = \frac{\text{min}_{1leqi<jleqn} d(i,j)}{\text{max}_{1\leq k\leq n} d'(k)}, \tag{15}$$

where $d(i,j)$ represents the distance between clusters i and j, and $d'(k)$ measures the intra-cluster distance of cluster k (10). We use the Euclidean distance to compute the distances. We seek a high Dunn index value, as this represents high intra-cluster similarity (or low values of $d'(k)$) and low inter-cluster similarity (or low values of $d(i,j)$). A clustering with a high Dunn index value will have very distinct clusters that are each tightly bound.

3.5.2 External Evaluation

We compute the following external evaluation methods to compare two clusterings. This allows the determination of whether or not two clusterings are similar or different.

The *Rand index* (RI) computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. One can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It is computed as

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{16}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives (38). One issue with the Rand index is that false positives and false negatives are equally weighted. This may be an undesirable characteristic for some clustering applications. The Rand index yields a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are exactly the same.

Hubert and Arabie's *adjusted Rand index* is the corrected-for-chance version of the Rand index (21). Though the Rand Index may only yield a value between 0 and 1, the adjusted Rand index can yield negative values if the index is less than the expected index. Given a set S of n elements, and two clusterings of these elements $X = \{X_1, X_2, ..., X_r\}$ and $Y = \{Y_1, Y_2, ..., Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$, where each entry n_{ij} denotes the number of objects in common between X_i and Y_j: $n_{ij} = |X_i \cap Y_j|$.

$X \backslash Y$	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{22}	...	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	a_r
X_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2	...	b_s	

Because the Rand index is highly dependent upon the number of clusters, Morey and Agresti propose a correction to the Rand statistic so that a measure of classification agreement would deal with different numbers of categories in each classification. It also corrects for expected agreement due to chance. This adjusted Rand statistic equals one when there is perfect agreement and equals zero when agreement is the same as change. Negative values indicate agreement less than that expected from chance alone (37; 13; 44).

Using the values in the contingency table, the adjusted Rand index (ARI) is calculated as

$$ARI = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}} \tag{17}$$

The Fowlkes-Mallows (FM) index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications (14). The higher the value of the Fowlkes-Mallows index the more similar the clusters and the benchmark classifications are. It can be computed using the following formula:

$$FM = \sqrt{\frac{TP}{TP+FP}\frac{TP}{TP+FN}}, \tag{18}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The Jaccard index is used to quantify the similarity between two datasets. It takes on a value between 0 and 1. An index of 1 means that the two dataset are identical, and an index of 0 indicates that the datasets have no common elements (22). The Jaccard index is defined by the following formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}, \tag{19}$$

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.

4 RESULTS

4.1 Introduction to Eye Tracking Data

Eye movement data, typically captured by eye tracking systems using infrared cameras to illuminate the eyes, provide gaze-informed insight into visual attention. We will use the *HMM with a multivariate normal probability distribution* to cluster raw eye movement data to inform distinct patterns of eye movement. By accounting for the sequential nature of the data (i.e., time-dependent samples), along with the two-dimensional spatial location of the data, we can distinguish eye movement patterns based on the direction, speed, and location of the patterns.

We will apply these methods to a subset of existing eye tracking data for two participants, whom we refer to as Participant A and Participant B, respectively, for the remainder of this article, who perform a task looking for four colored dots overlaid on a synthetic aperture radar (SAR) image in a set order [1]. See Figure 1 for an example of a zoomed-in section of the image. The investigation will use visualizations to determine similarities and differences between the two participants' eye movement behaviors. The two participants' datasets will be labeled Participant A and Participant B, respectively.

[1] This data set is part of a larger effort to develop and validate an algorithm to go from raw eye movement data to meaningful content in a dynamic, user-driven environment without the need for hand coding. Here, we are focusing on the first task only (out of the four tasks used in the larger study). Sixteen participants completed the study (9).

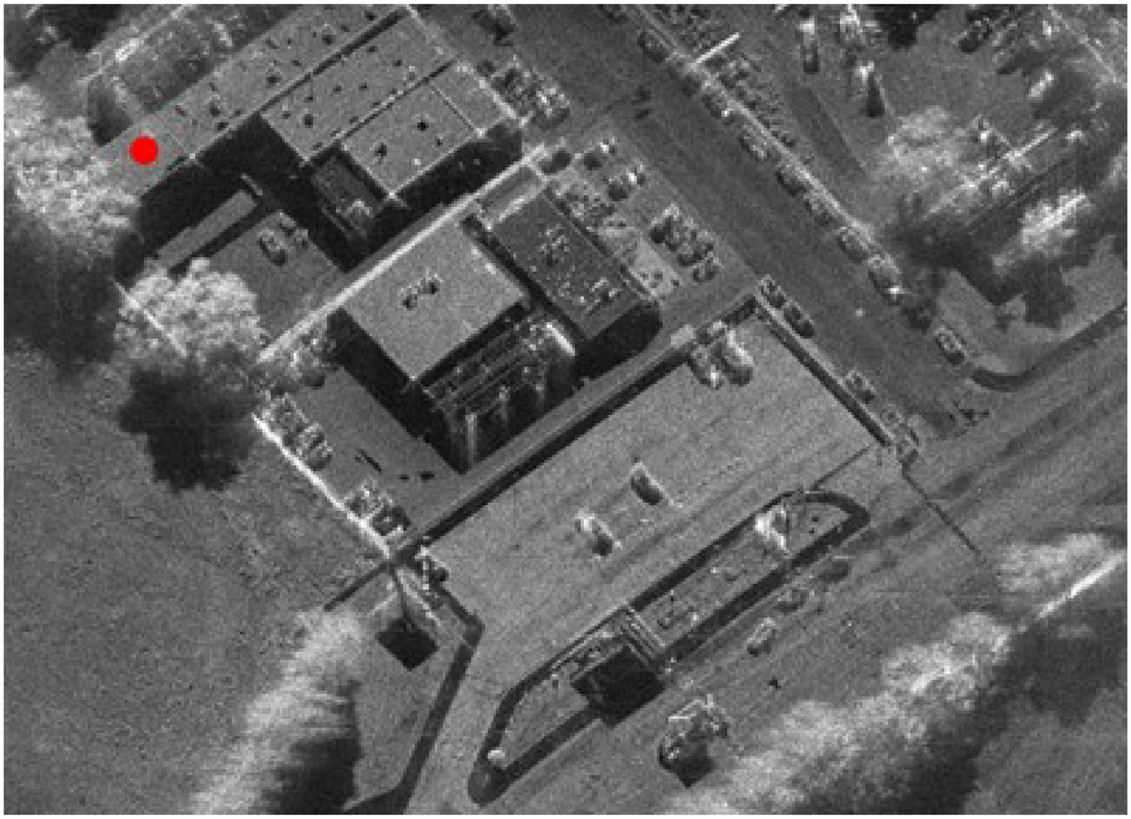


Figure 1. Example of zoomed-in section of SAR image. Participants panned through the image, finding four colored dots in a set order. (Image UUR)

The spatial (x,y) coordinates for the four targets that are to be found in succession (labeled as targets one through four) are as follows [2]

- Target One: (1.457,54.198)
- Target Two: (54.787,21.792)
- Target Three: (20.437,2.311)
- Target Four: (71.439,81.370)

[2] We note that in the dataset, the horizontal scale goes from zero to 80 from left to right. However, the vertical scale goes from zero to 100 from *top to bottom*. Thus, the origin has coordinates (0,100).

4.2 Features of Eye Movement Data as Covariates

We compute the following derived features of the eye movement data that can be incorporated as covariates in the HMM in the hopes of identifying patterns in the data and being able to distinguish one entire eye movement pattern from another.

- **Length Ratio (lenratio)**: the ratio of the total Euclidean distance traveled from one target to another to the straight-line distance from one target to another. The length ratio helps to determine how circuitous of a path the participant takes to go from one target to another. In a trial, there can be up to four distinct values of the length ratio. Those four values can be measured from the starting point of the trial to target one, target one to target two, target two to target three, and target three to target four. Thus, for each data point that is within these ranges, the value of the length ratio will be the same.
- **Angle**: the angle is the direction of movement computed from the previous data point. There can potentially be a distinct value for the angle at each data point.
- **Angle Difference (anglediffs)**: the difference in the angle at the current data point versus the previous data point. This helps to measure the change in direction of the eye movement from point to point. There can potentially be a distinct value for the angle difference at each data point.
- **Total Angles (totalAngles)**: this measure is the cumulative angle measures for the entire trial. It encompasses all of the changes in the direction the participant makes during the entire trial. This measure is a function of the cumulative angle differences measured from target to target. Therefore, there can be up to four distinct values for the total angles, with each data point within the ranges between targets having the same value for the total angles.

4.3 Results for Participant A's Eye Tracking Data

Figure 2 contains a plot of the path of person one's eye movement data. The color lines on the plot indicate the following:

- Starting point to target one: Black
- Target one to target two: Blue
- Target two to target three: Orange
- Target three to target four: Red

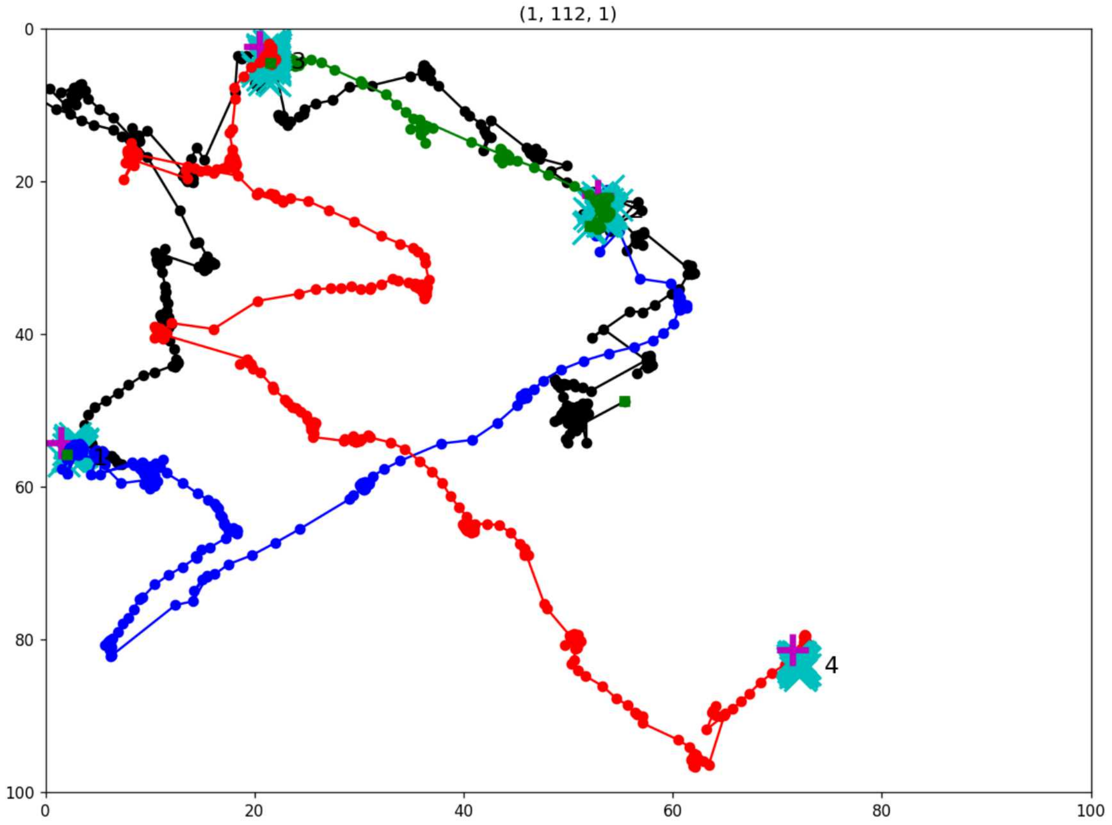


Figure 2. Plot of Eye Movement Data for Participant A

We wish to use the HMM to further distinguish eye movement patterns in the spatial locations of the data for Participant A.

4.3.1 Implementation of Existing GMM and Visualization Methods

The genesis of this project is the shortcomings in existing GMM and visualization methods. Before we discuss the HMM implementation results and the implementation of our developed visualization tools, we first discuss the results from fitting the GMM implementations and visualizations of (15), which assume that each data point is i.i.d. and follows a certain covariance matrix structure, and (33), which assume that each data vector is independent but the elements of the vector follow an autoregressive time series model.

When fitting a GMM, we assume that each data vector $\mathbf{y}$ follows the probability density function

$$f(\mathbf{y}|\vartheta) = \sum_{g=1}^{G} \pi_g \frac{\exp\{-\frac{1}{2}(\mathbf{y}_i - \mu_g)^T \Sigma_g^{-1} (\mathbf{y}_i - \mu_g)\}}{\sqrt{\det(2\pi\Sigma_g)}}, \tag{20}$$

where μ_g is the mean vector and Σ_g is the covariance matrix of component g. The resulting complete-data likelihood is

$$\mathcal{L}_C(\pi_g, \mu_g, \Sigma_g) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f(x_i|\mu_g, \Sigma)]^{z_{ig}}, \tag{21}$$

where z_{ig} denotes the membership of observation i in component g so that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise. An expectation-maximization (EM) algorithm (8) is implemented to estimate all of the parameters in the model. Afterwards, we can estimate the maximum likelihood estimator (MLE) of the classification of each data point as $\{j|z^*_{ij} = \max_g z^*_{ig}\}$, and the classification uncertainty is computed as $(1 - \max_g z^*_{ig})$. In the implementation of (15) for the i.i.d. case, geometric cross-cluster constraints in multivariate normal mixtures are taken into account by parameterizing covariance matrices through an eigenvalue decomposition in the form

$$\Sigma_g = \lambda_g D_g A_g D_g^T, \tag{22}$$

where D_g is the orthogonal matrix of eigenvectors, A_g is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_g is an associated constant of proportionality.

Below in Figure 3(b) is a clustering and uncertainty plot for Participant A's data using the mclust package, the associated R package with (15). In this plot, the clustering uncertainty ellipses drawn from the mclust package do not match up well with the observed data because it does not factor in the temporal correlation between observations. The plot indicates this clustering method is not appropriate for our data.

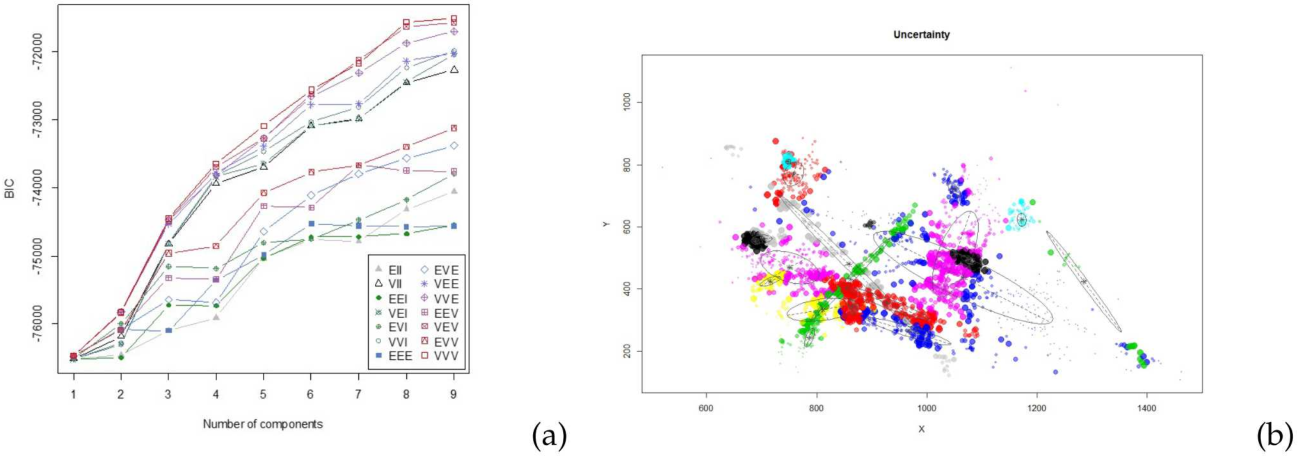


Figure 3. BIC and clustering results and uncertainty plots for R package mclust applied to eyetracking data. The model chosen by the highest BIC value (a) is a model with 20 clusters and parametrization VVV. (b) consists of the clustering results and associated uncertainty, which is represented by the ellipses.

(33) assumes that each vector is independent and consists of temporally-correlated scalar observations. Let Y_t denote the observation at time t. The temporal correlation between observations is accounted by the modified Cholesky decomposition of the inverse covariance matrix,

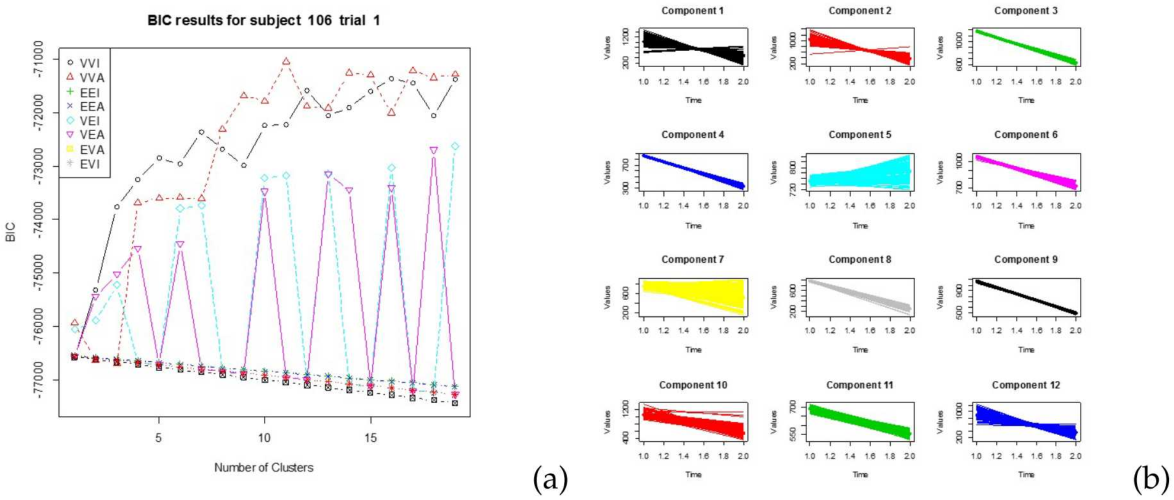
$$\Sigma^{-1} = T' D^{-1} T,$$

where T is a unique lower triangular matrix with diagonal elements 1 and D is a unique diagonal matrix with strictly positive diagonal entries. The values of T and D have interpretations as generalized autoregressive parameters and innovation variances, respectively, so that the linear least-squares predictor of Y_t, based on $Y_{t-1}, ..., Y_1$, can be written as

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1} (-\phi_{ts})(Y_s - \mu_s) + \sqrt{d_t}\epsilon_t, \tag{23}$$

where $\epsilon_t \sim N(0, 1)$, the ϕ_{ts} are the (sub-diagonal) elements of T and the d_t are the diagonal elements of D.

Below in Figure 4 are the plots currently available in the longclust package, the associated R package for (33), for longitudinal data applied to Participant A's data. The R package longclust currently does not have the capability to plot the clustering results and uncertainty for fitting a GMM to longitudinal multivariate data that the mclust package does for i.i.d. data. It is unclear what the values in the time plots in Figure 4(b) represent. Furthermore, we cannot visualize the clustering and uncertainty results.



(a) (b)

Figure 4. Currently available plots for R package longclust applied to eyetracking data. The model chosen by the highest BIC value (a) is a model with 12 clusters. (b) consists of time plots for the 12 clusters and appears to be the values for a parameter associated with the 12 clusters over the running of the EM algorithm until convergence. However, it is unclear what that parameter is.

4.3.2 HMM Clustering Results with No Covariates

To find these patterns, we fit a HMM to the spatial locations of Participant A's eye movement data. We do not consider any features of eye movement data that can be used as covariates to better describe Participant A's eye movement patterns. Using the Bayesian information criterion (BIC), we find the optimal HMM is one with 12 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant A's first trial data. In all of the visualizations, the four targets are indicated by larger percent signs with the following colors:

- Target One: Red
- Target Two: Yellow
- Target Three: Green
- Target Four: Blue

Figure 5 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). In these plots, the four targets are indicated by the larger colored percent signs. The ellipses seen in Figure 5(b) indicate that most of the clusters are pretty tightly constrained and non-overlapping, which shows that the 12 clusters assigned by the HMM are mostly distinct eye movement patterns, when the type of eye movement pattern (fixating around a target or moving across the image, for examples) and the location of the pattern are considered together.

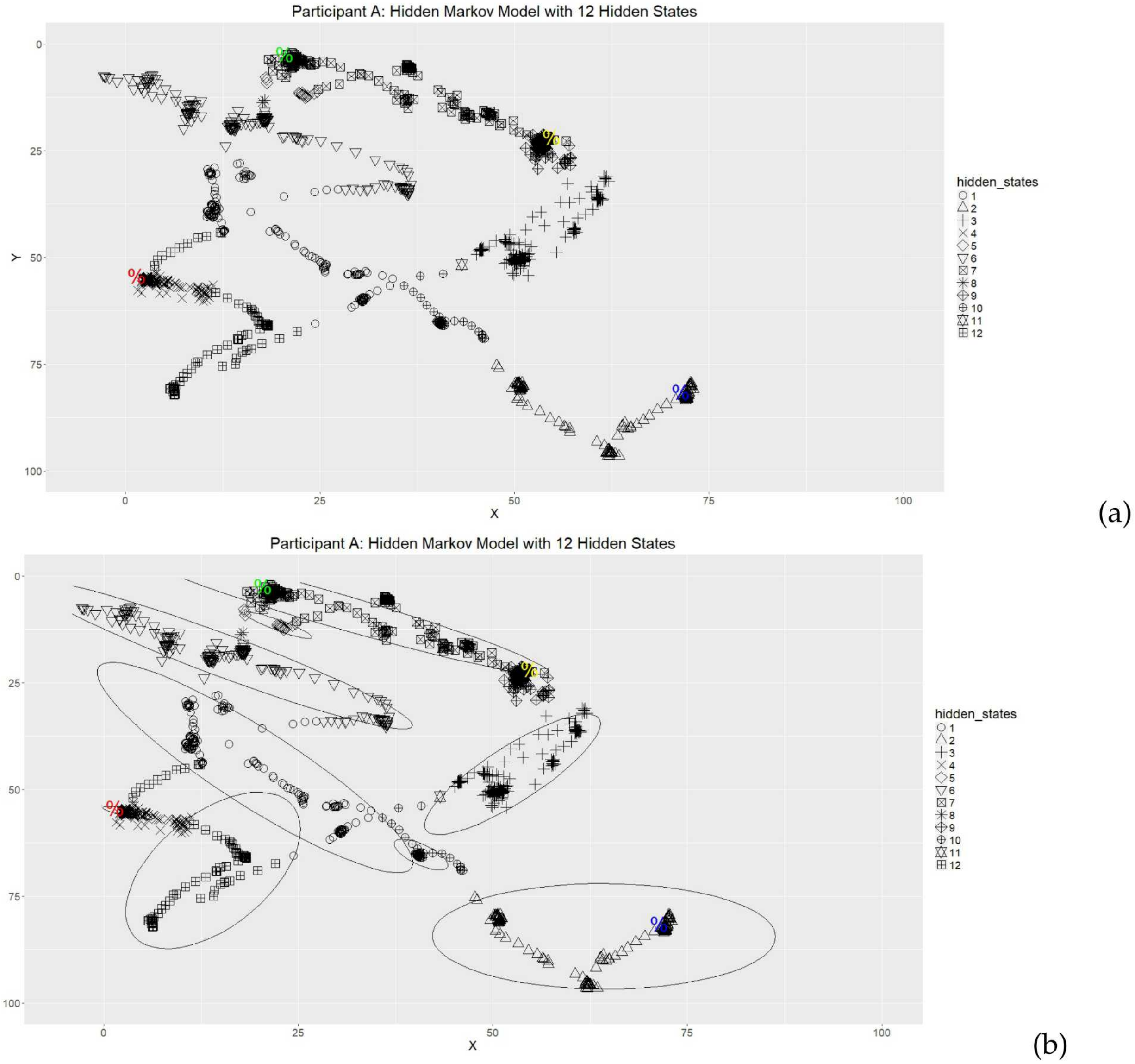


Figure 5. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant A

Figure 6 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 6(b) that clusters two, three, six, and 12 (as indicated by their respective symbols in the legend) all have points classified to these clusters with low uncertainty, as indicated by the solidly dark blue colors. On the other hand, clusters one, five, nine, and 10 (as indicated by their respective symbols in the legend) all have points classified with relatively high uncertainty, as indicated by the lighter blue colors for these data points. While we cannot definitively say what are the specific eye movement patterns to each specific cluster, as probabilistic clustering models are unsupervised methods, these results indicate that clusters two, three, six, and 12 correspond to distinct eye movement patterns. While the HMM seem to want to distinguish eye movement patterns with clusters one, five, nine, and 10, the level of distinction is not as clear.

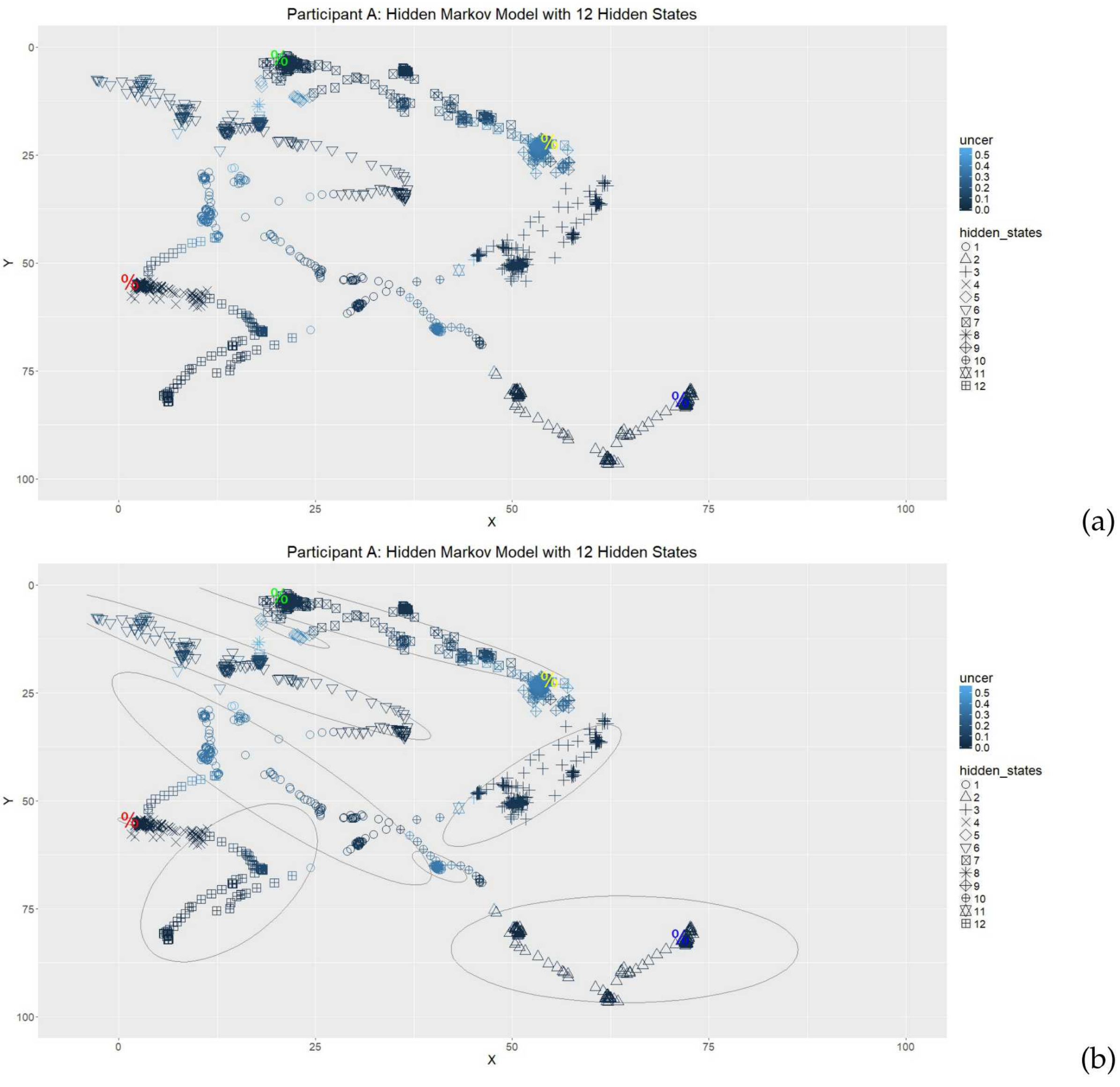


Figure 6. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant A

Figure 7 contains a time plot of the changes in cluster assignments over time as the trial progresses. We see certain clusters, such as two, four, seven, and nine, where the participant's eyes spend a relatively longer amount of time. These correspond to behaviors where the eyes are fixating around the targets. On the other hand, there are clusters such as five and eight, that the participant hardly spends any time. These seems to indicate very fleeting behavior.

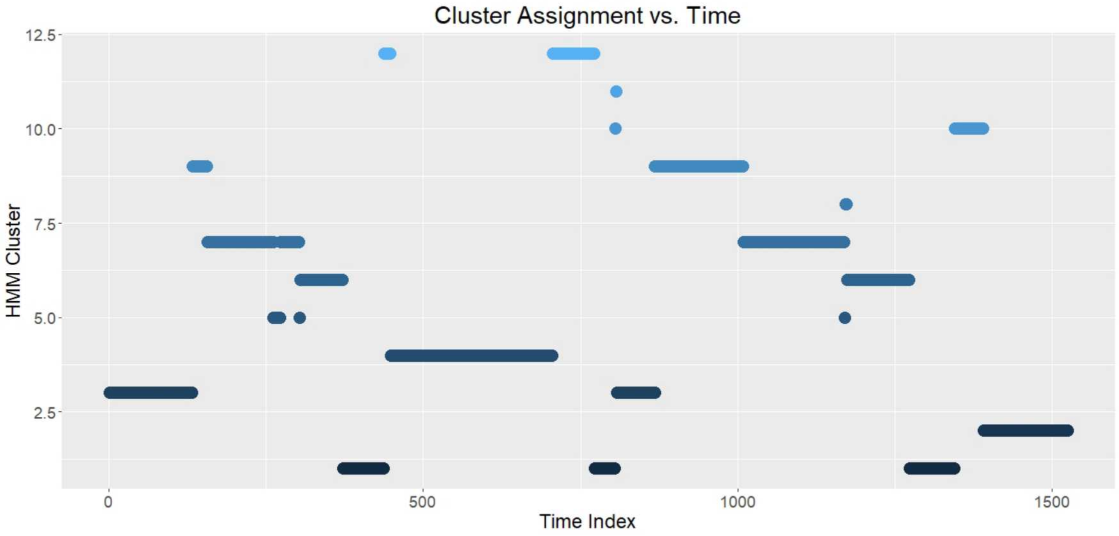


Figure 7. Time Plot of Cluster Assignments for Participant A

4.3.3 HMM Clustering Results with Covariates lenratio, angle, anglediffs, and totalAngles

We compare the HMM clustering results for Participant A when we incorporate covariates into the clustering. We fit a HMM to the spatial locations of Participant A's eye movement data with the length ratio, angle, angle difference, and total angles as covariates. Using the Bayesian information criterion (BIC), we find the optimal HMM is one with 10 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant A.

Figure 8 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). In these plots, the four targets are indicated by the larger colored percent signs. The ellipses seen in Figure 8(b) indicate that a lot of the clusters overlap with one another and are wider compared to the clusters for when no covariates are considered. This indicates that the added covariates do not improve the clustering results.

Figure 9 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 9(b) that most of the points have been clustered to their assigned clusters with relatively low uncertainty, as indicated by the fact that most data points are marked with dark blue colors. However, the ellipses show the great amount of overlap between clusters, which shows that even though the data points are clustered with high confidence, the quality of the clustering is very questionable due to the amount of overlap between clusters.

Figure 10 contains a time plot of the changes in cluster assignments over time as the trial progresses. We see certain clusters, such as one, five, seven, and nine, where the participant's eyes spend a relatively longer amount of time. These correspond to behaviors where the eyes are fixating around the targets. On the other hand, there are clusters such as seven and eight, that are revisited multiple times that the participant spends relatively small amounts of time in on a given visit These seem to indicate a panning behavior where the participant's eyes are moving quickly through one part of the image to another.

4.3.4 HMM Clustering Results with Covariate angle

A time series regression analysis of the spatial location of person's eye movement data as the response variable and the length ratio, angle, angle differences, and total angles as covariates reveals that the angle is the most dominant covariate. Therefore, we fit a HMM to the eye movement data with angle as the lone covariate.

Using the Bayesian information criterion (BIC), we find the optimal HMM is one with

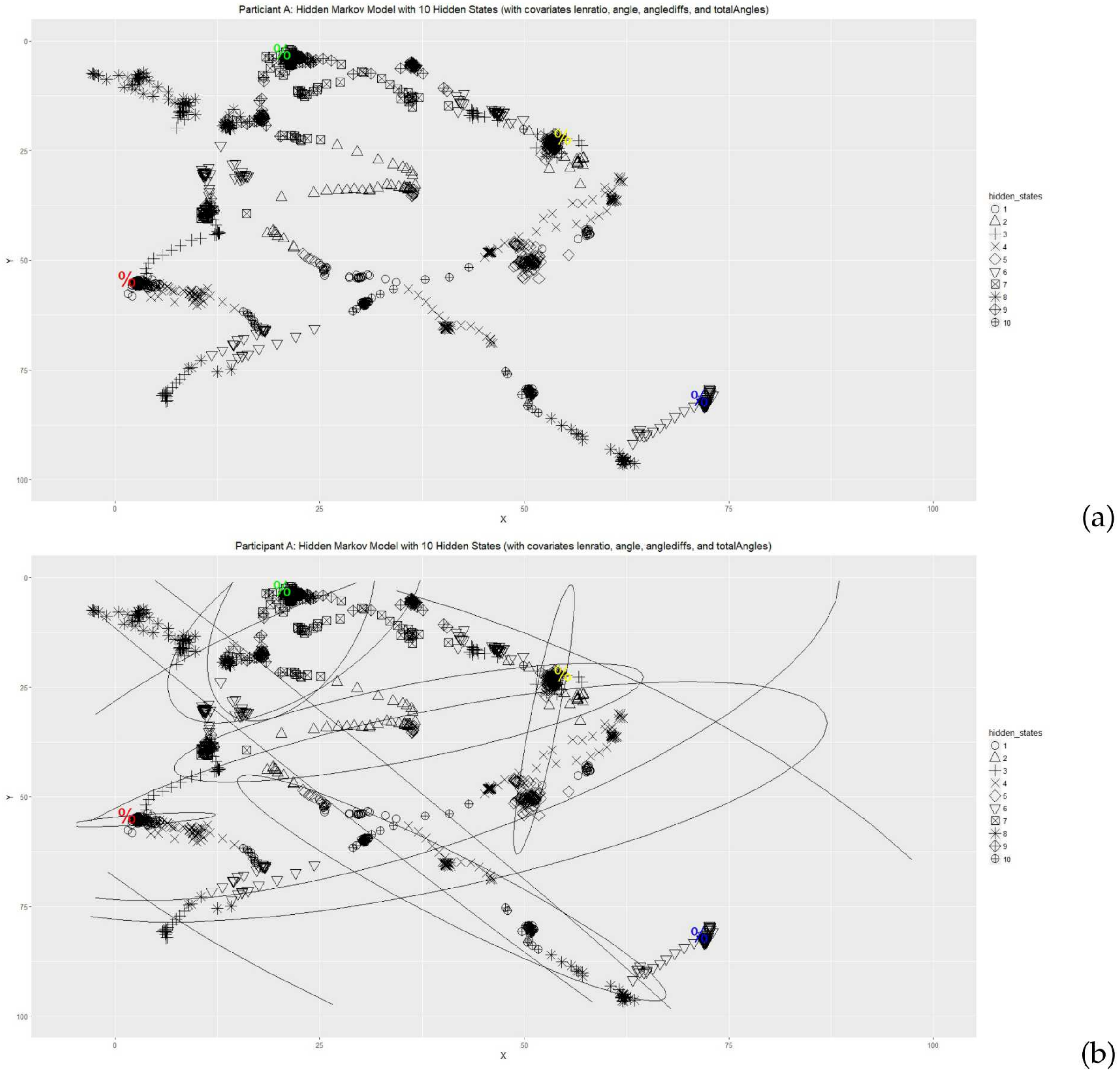
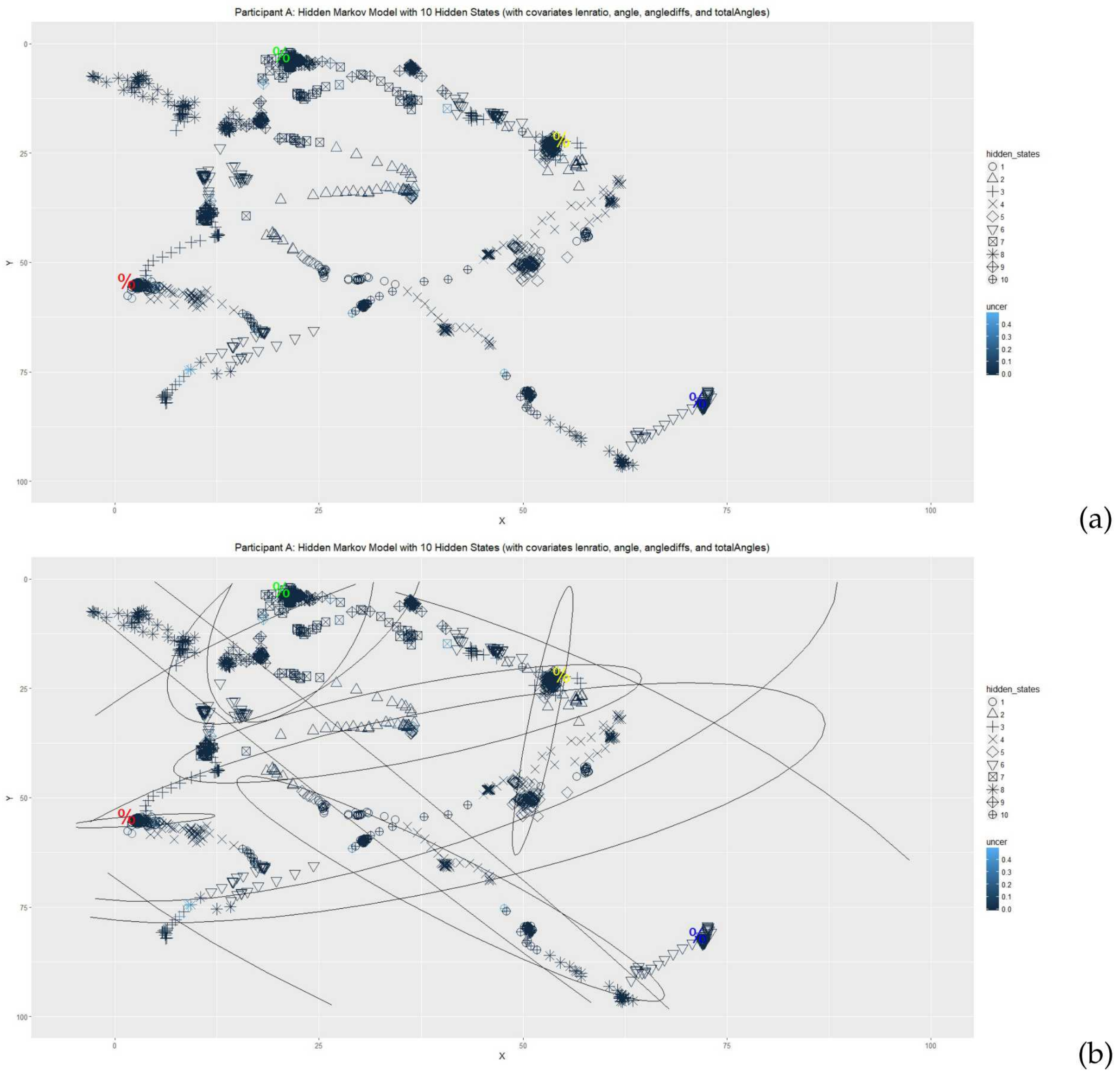


Figure 8. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant A with covariates lenratio, angle, anglediffs, and totalAngles

12 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant A's data data.

Figure 11 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). In these plots, the four targets are indicated by the larger


Participant A: Hidden Markov Model with 10 Hidden States (with covariates lenratio, angle, anglediffs, and totalAngles)

(a)

Participant A: Hidden Markov Model with 10 Hidden States (with covariates lenratio, angle, anglediffs, and totalAngles)

(b)

Figure 9. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant A with covariates lenratio, angle, anglediffs, and totalAngles

colored percent signs. The ellipses seen in Figure 11(b) indicate that there are much fewer overlapping clusters as compared to when multiple covariates are incorporated. This indicates that distinct patterns in the eye movement data can be identified better by the HMM. However, the clusters are not quite as distinct as when no covariates are included.

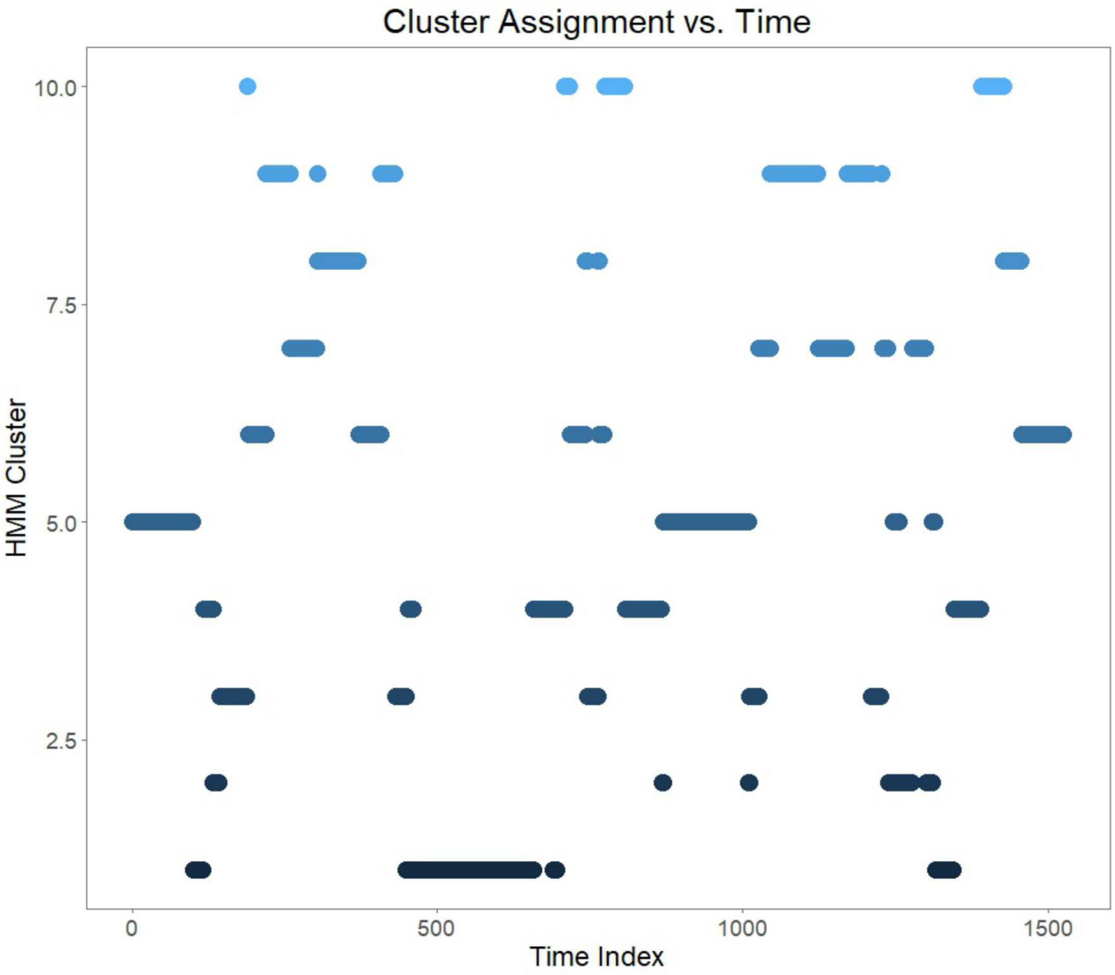


Figure 10. Time Plot of Cluster Assignments for Participant A with covariates lenratio, angle, anglediffs, and totalAngles

Figure 12 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 12(b) that most of the points have been clustered to their assigned clusters with relatively low uncertainty, as indicated by the fact that most data points are marked with dark blue colors. The very evident light blue points are largely concentrated around the center of the image (around (x,y) coordinates (50,50)), with most of the data points clustered into cluster 12. While there is still some overlap between the ellipses, the clusters are much more distinct when angle is the only covariate considered, as opposed to when many more covariates are considered. This is likely due to the fact that the angle is part of the polar coordinate representation of the (x,y) coordinate system (r, θ). Therefore, the angle is more directly tied into the spatial location, and the change in direction is more significant determinant of eye movement

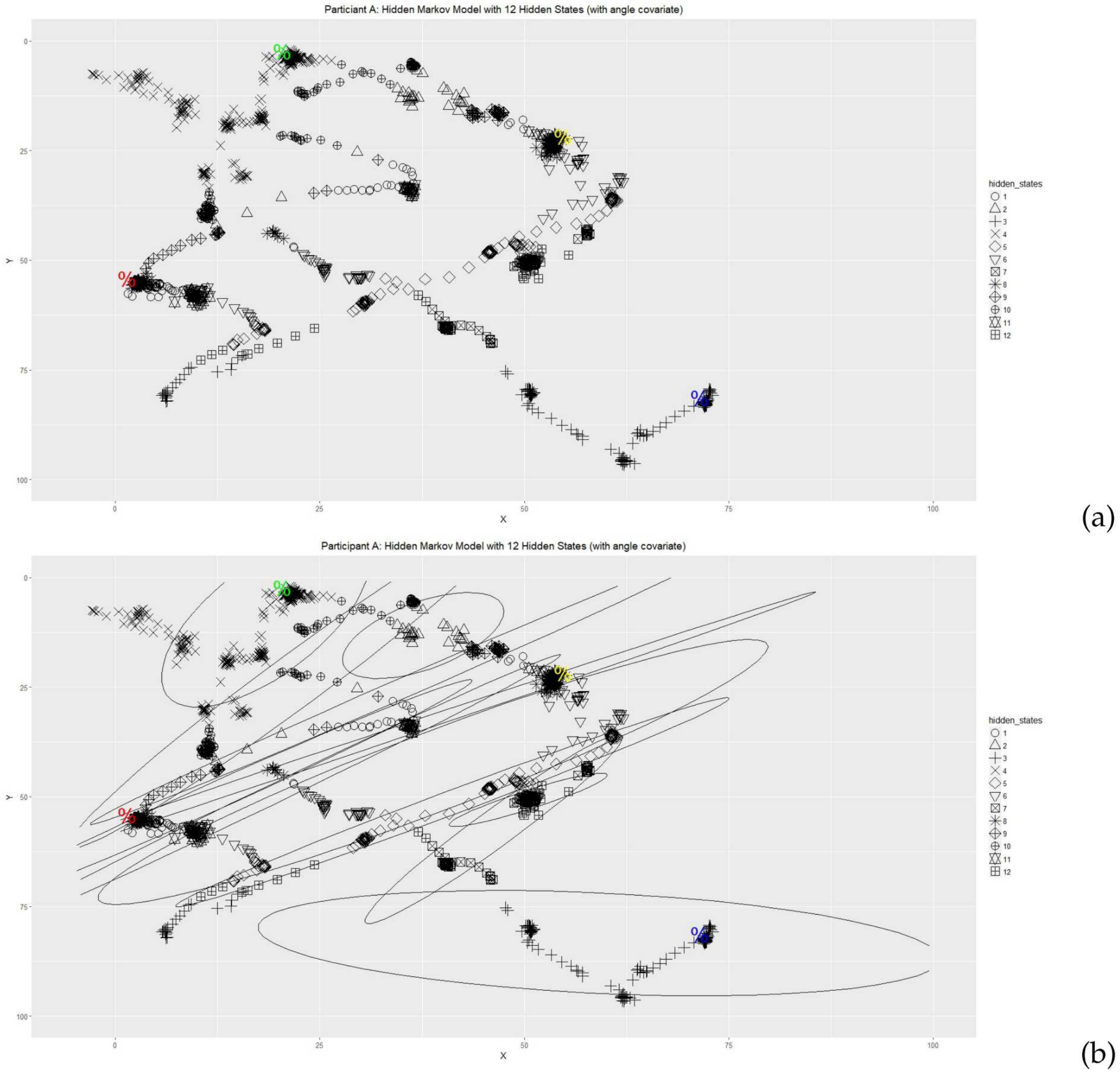


Figure 11. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant A with covariate angle

patterns.

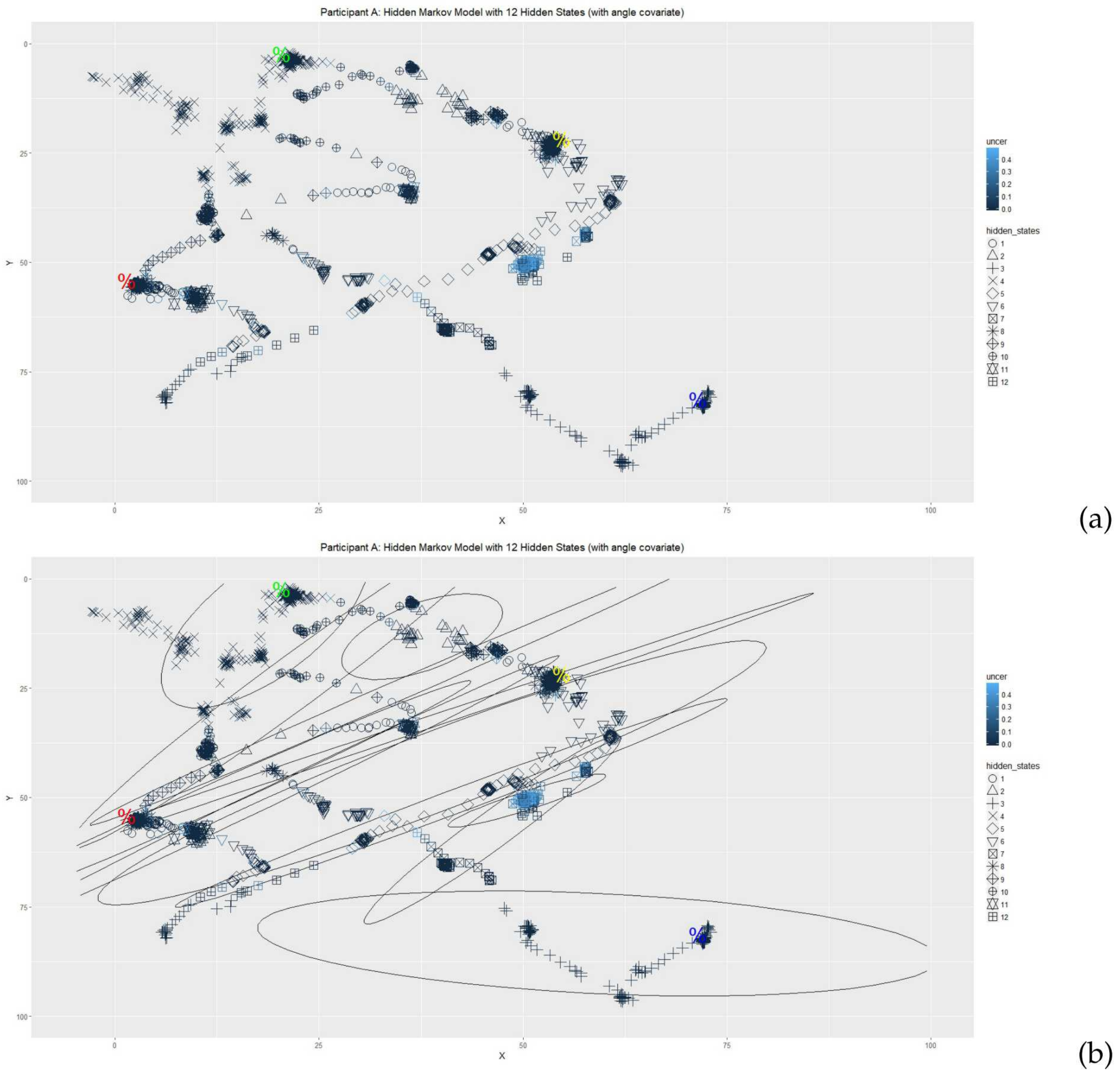


Figure 12. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant A with covariate angle

Figure 13 contains a time plot of the changes in cluster assignments over time as the trial progresses. We see certain clusters, such as three, four, and eight, where the participant's eyes spend a relatively longer amount of time. Cluster 12 has a longer stretch at the beginning of the trial, which indicates where participant A starts. These correspond to behaviors where the eyes are fixating around the targets and the participant starts the trial. On the other hand, there are clusters such as one, two, six, and nine, that are revisited multiple times that the participant hardly spends any time. These seem to indicate mostly panning behavior where the participant's eyes are moving quickly through one part of the image to another.

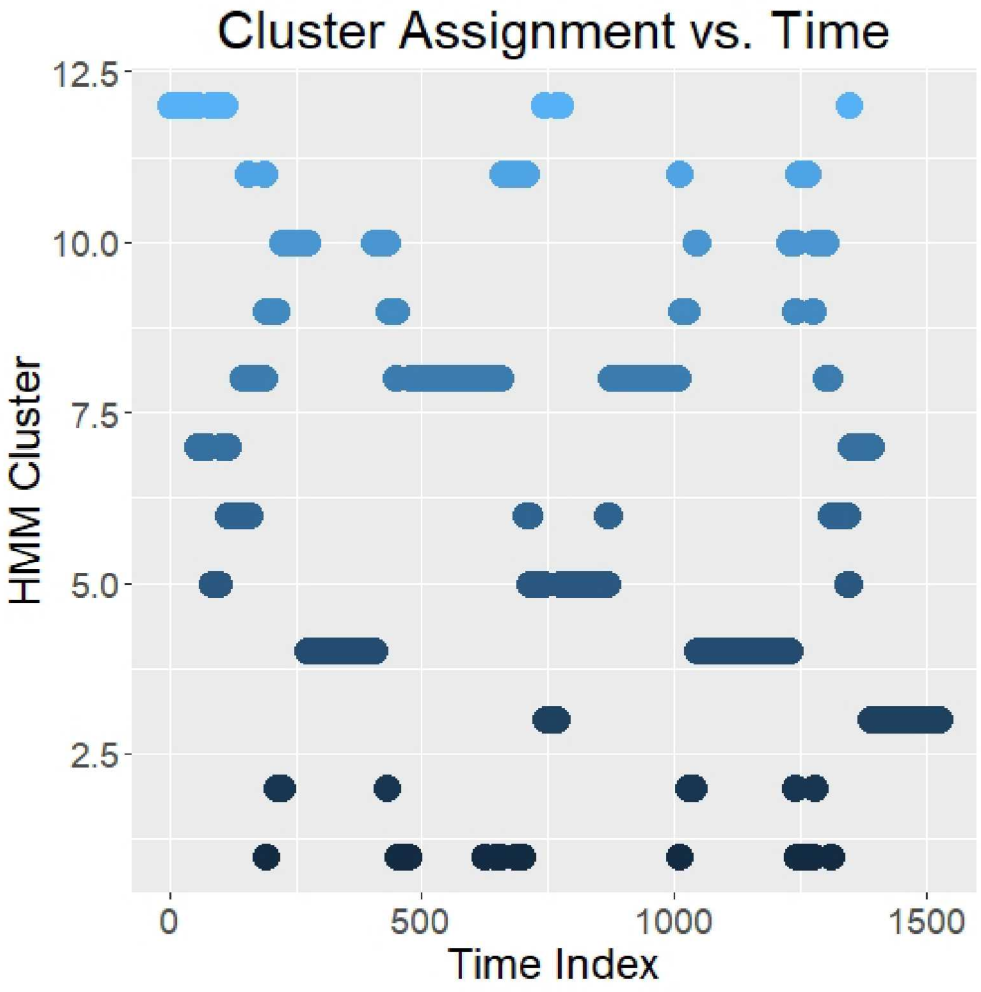


Figure 13. Time Plot of Cluster Assignments for Participant A with covariate angle

4.3.5 Pairwise Comparisons of Clusterings

We compute the Dunn index to look at the separability of clusters within each clustering done above. We also compute the five external measures to do pairwise comparisons of the clusterings. The external measures will use the following shorthands in parentheses in the list below:

- Rand index (Rand)
- Hubert and Arabie's adjusted Rand index (HA)
- Morey and Agresti's adjusted Rand index (MA)
- Fowlkes and Mallows' index (FM)
- Jaccard index (Jaccard)

For the sake of brevity, in both tables, we refer to "Multiple" covariates below as the model including the covariates length ratio, angle, angle differences, and total angle differences. Table 1 contains the Dunn index values for Participant A's HMM modelings.

Covariates	Dunn index
None	0.000750678
Angle	0.001118827
Multiple	0.0001174093

Table 1. Dunn Index Values Measuring Separability of Clusters for Participant A's HMM Modelings

Since higher Dunn index values indicate more separability between clusters, the model with the angle covariate has the highest separability, followed by the model with no covariates. The Dunn index values support our findings in the visualizations that the model with multiple covariates had much more overlap between clusters. However, from the visualizations, it is not clear that the model with the angle covariate has more separability between clusters than the model with no covariates. The visualizations reveal the specific nature of the separability between clusters (i.e. which clusters are more distinct than each other), which would not be revealed in the Dunn index, a global measure over the entire clustering.

The results for the external measures in Table 2 indicate that the models with no covariates and angle covariate have the most similarity, which support our visual findings.

Table 2. Pairwise Comparisons of Similarities Between Participant A's HMM Modelings.

Model 1 Covariates	Model 2 Covariates	Rand	HA	MA	FM	Jaccard
None	Angle	0.8589639	0.3646231	0.3674708	0.4461954	0.2860429
None	Multiple	0.8647726	0.3460608	0.3492777	0.4226430	0.2679437
Angle	Multiple	0.8402599	0.2802286	0.2834553	0.3722151	0.2278067

4.4 Results for Participant B's Eye Tracking Data

As an additional point of illustration and a comparison to the analysis done for Participant A, we wish to distinguish eye movement patterns in the spatial locations of the data for an additional participant, in order to see whether similar patterns hold or those seen for Participant A were due to idiosyncrasies in his/her data. Figure 14 is a plot of the path of Participant B's data.

Just as seen in Figure 2 for participant A, the color lines on the plot indicate the following:

- Starting point to target one: Black
- Target one to target two: Blue
- Target two to target three: Orange
- Target three to target four: Red

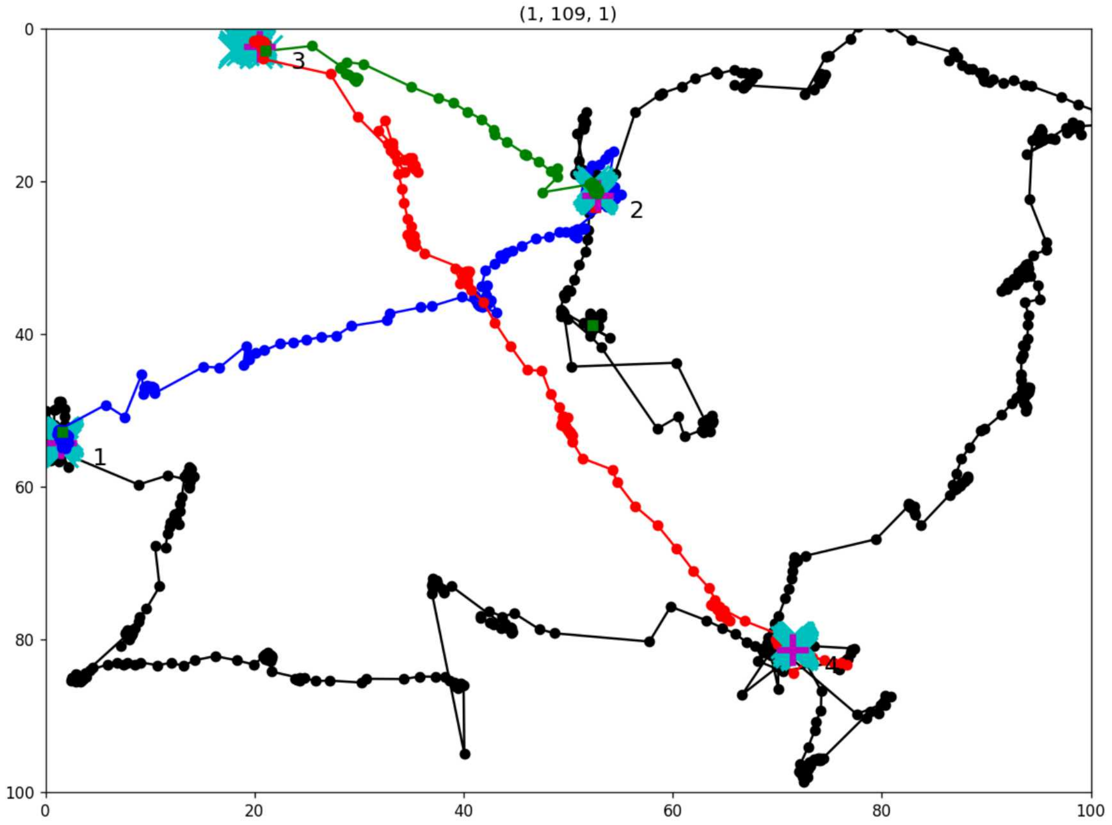


Figure 14. Plot of Eye Movement Data for Participant B

4.4.1 HMM Clustering Results with No Covariates

To find these patterns, we fit a HMM to the spatial locations of Participant B's eye movement data. We do not consider any features of eye movement data that can be used as covariates to better describe Participant B's eye movement patterns. Using the Bayesian information criterion (BIC), we find the optimal HMM is one with 14 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant B's data. In all of the visualizations, the four targets are indicated by larger percent signs with the following colors:

- Target One: Red
- Target Two: Yellow
- Target Three: Green
- Target Four: Blue

Figure 15 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). In these plots, the four targets are indicated by the larger colored percent signs. The ellipses seen in Figure 15 (b) indicate that most of the clusters are largely non-overlapping (with the notable exceptions of clusters three and four and clusters 11 and 14). This shows that most of the 14 clusters assigned by the HMM are mostly distinct eye movement patterns. But the two sets of overlapping clusters (three and four and 11 and 14) need to be investigated further as to what the patterns these clusters represent.

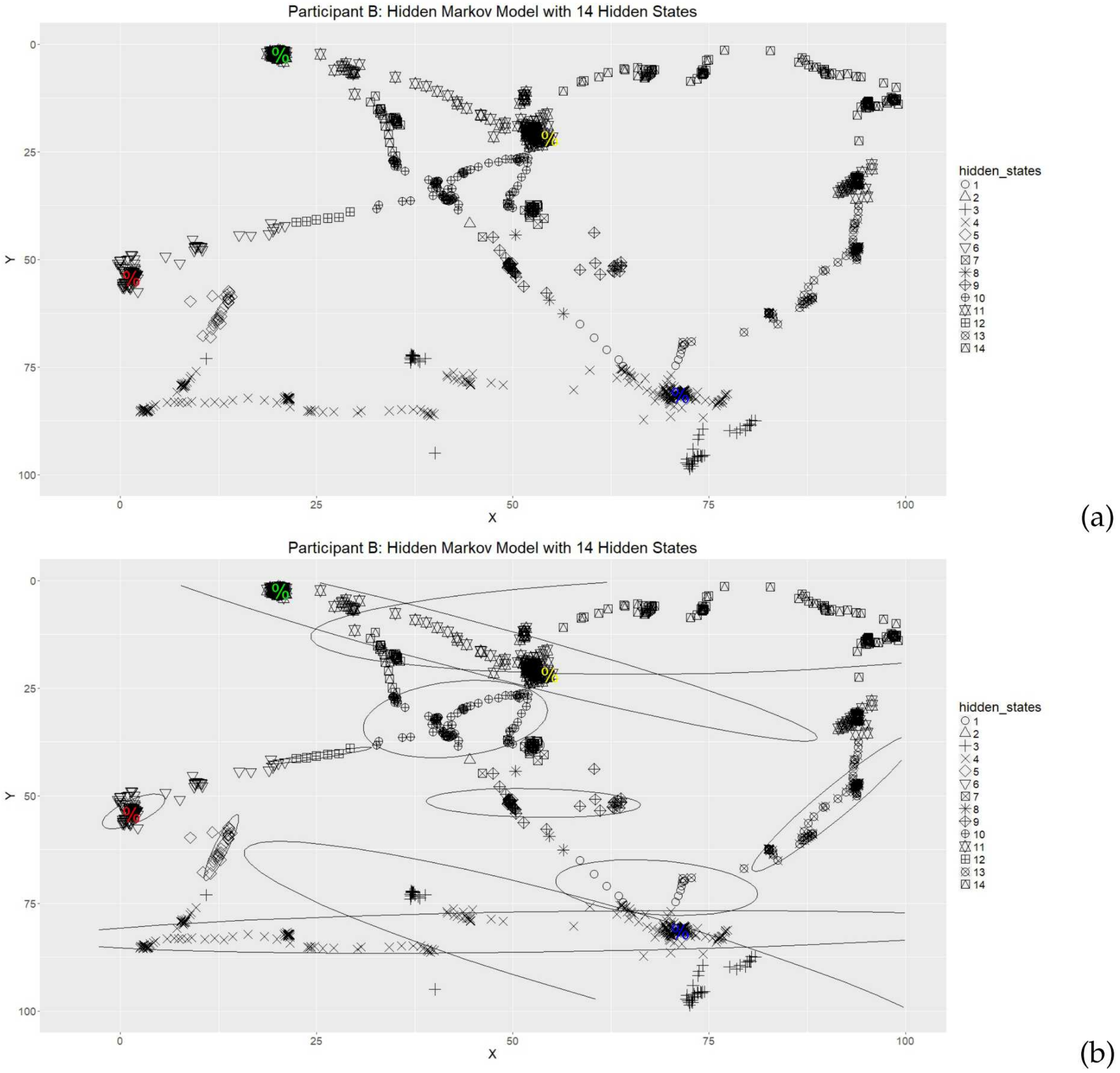


Figure 15. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant B

Figure 16 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 16(b) that most of the points with higher classification uncertainties (as indicated by the lighter blue colors) are around (x,y) coordinates (50,50) in the center of the image. With the variety of symbols in that part of the image, we see there is a lot more uncertainty with the classification of those data points. We see that particularly for clusters three, 10, and 14, the points on the outer edges of those clusters have higher uncertainties.

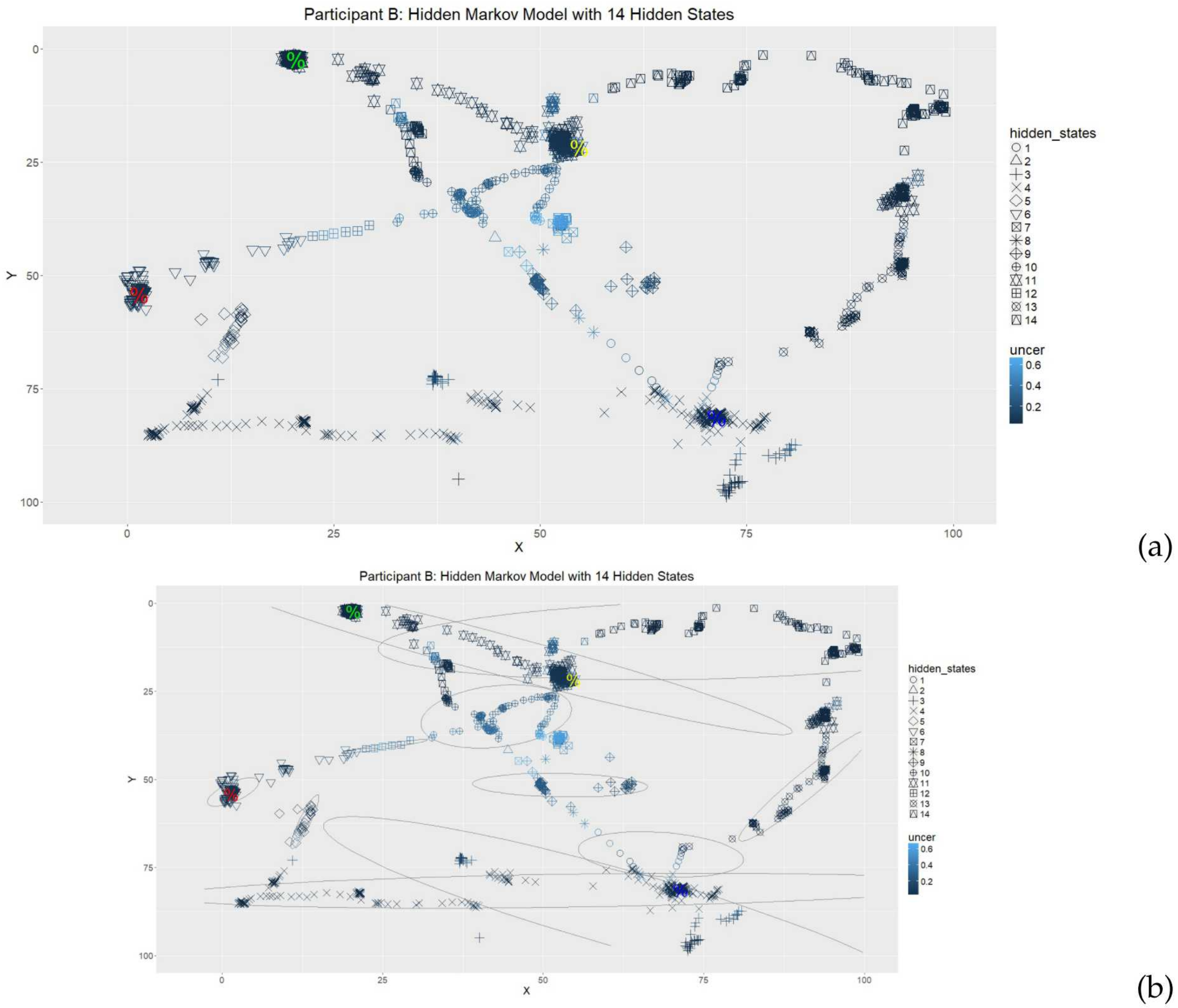


Figure 16. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant B

Figure 17 contains a time plot of the changes in cluster assignments over time as the trial progresses. We identify clusters four, six, 11, and 14 as those where the participant's eyes spend a relatively longer amount of time. These correspond to behaviors where the eyes are searching for and fixating around the four targets. For most of the other clusters, we don't see as many revisiting of clusters.

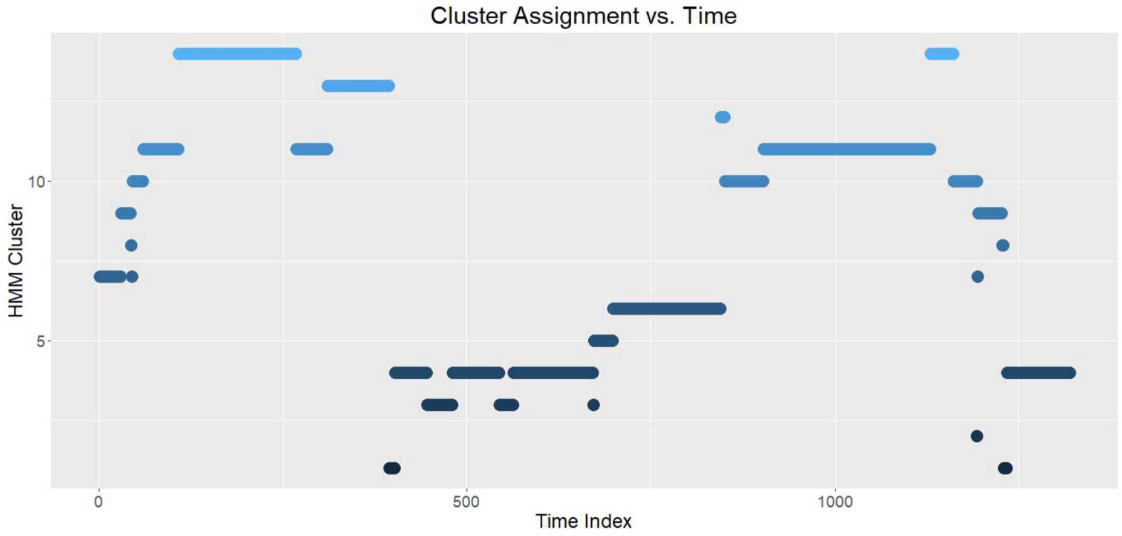


Figure 17. Time Plot of Cluster Assignments for Participant B

4.4.2 HMM Clustering Results with Covariates lenratio, angle, anglediffs, and totalAngles

We now compare the HMM clustering results for Participant B when we incorporate covariates into the clustering. We fit a HMM to the spatial locations of Participant B's eye movement data with the length ratio, angle, angle difference, and total angles as covariates. Using the Bayesian information criterion (BIC), we find the optimal HMM is one with 16 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant B's data.

Figure 18 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). The ellipses seen in Figure 18(b) indicate that a lot of the clusters overlap with one another and are wider compared to the clusters for when no covariates are considered. This indicates that the added covariates do not improve the clustering results. However, we do notice some similarities between the clusterings with no covariates. Both models have pretty horizontal cluster along the bottom part of the image (around y=83 and spanning $x \in (0, 75)$) that is fairly tight (corresponding to cluster 4 in both models) and a tight clustering across the right ride of the image (spanning $x \in (75, 100)$ and $y \in (20, 100)$) (corresponding to cluster 13 in the model with no covariates and cluster 3 in the model with multiple covariates).

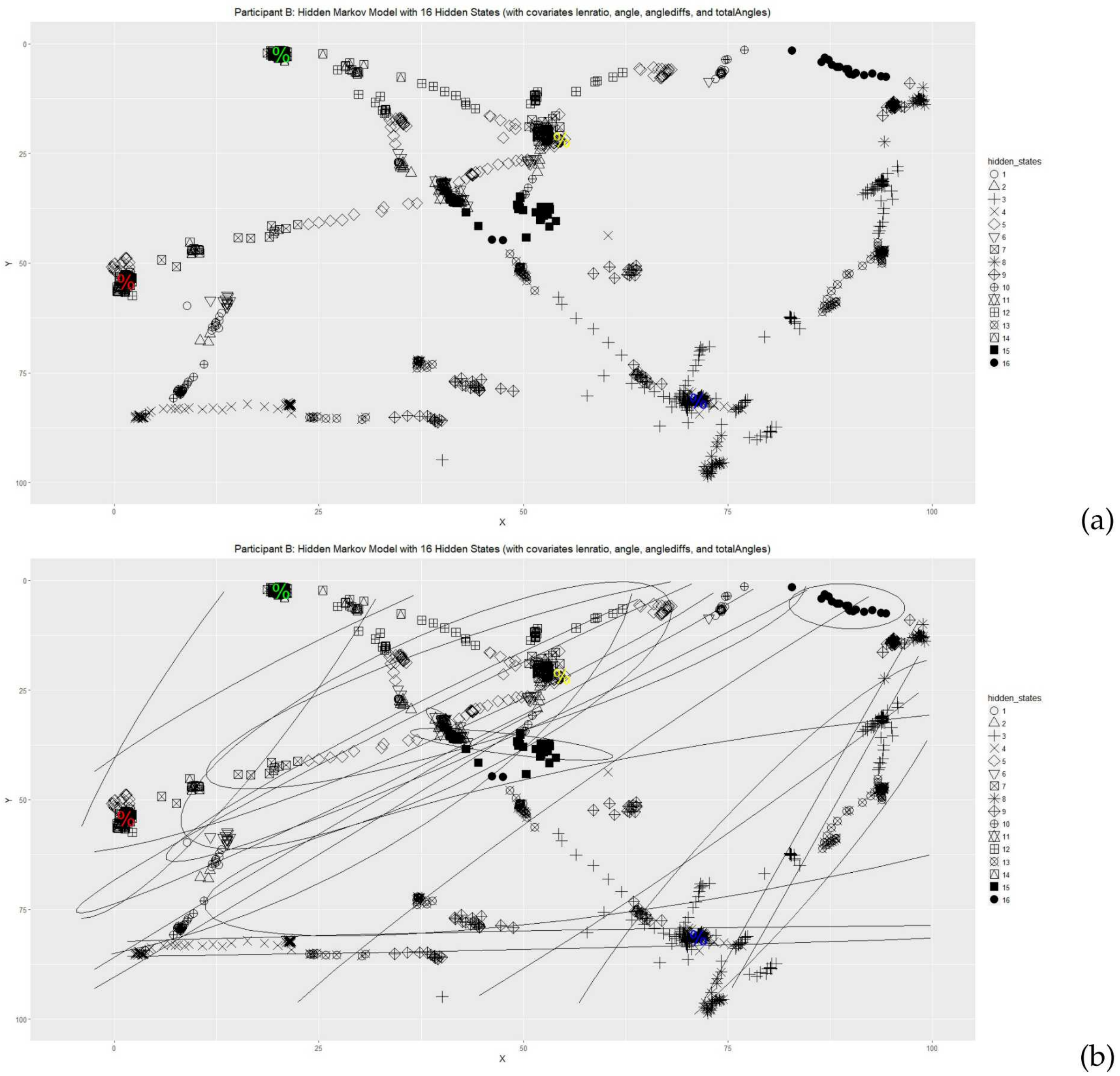


Figure 18. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant B with covariates lenratio, angle, anglediffs, and totalAngles

Figure 19 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 19(b) that most of the points have been clustered to their assigned clusters with relatively low uncertainty, as indicated by the fact that most data points are marked with dark blue colors. There are some points with higher uncertainty, as indicated by the lighter blue colors, around (x,y) coordinates

(40,30) in the image. However, the ellipses show the great amount of overlap between clusters, which shows that even though the data points are clustered with high confidence, the quality of the clustering is very questionable due to the amount of overlap between clusters.

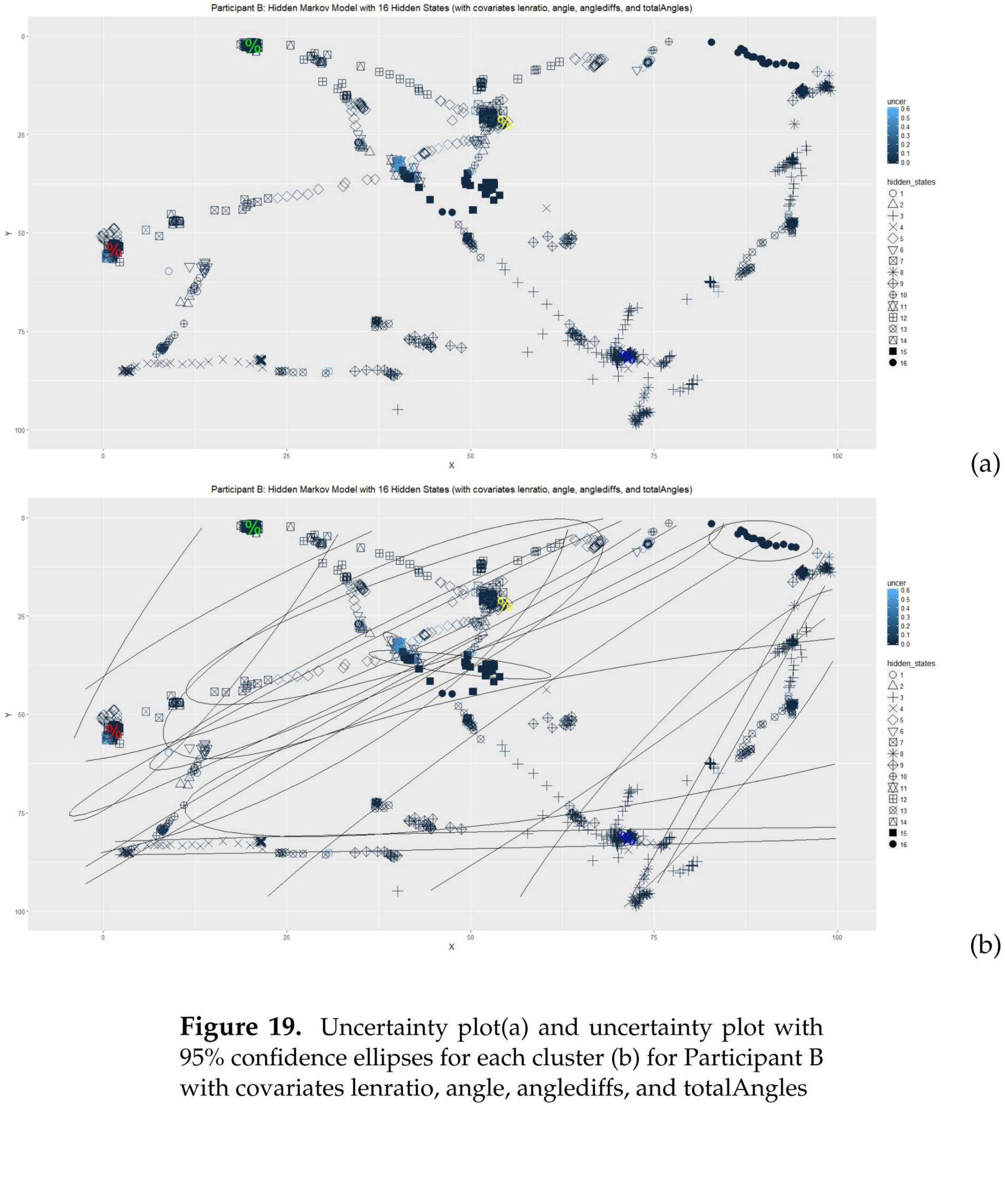


Figure 19. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant B with covariates lenratio, angle, anglediffs, and totalAngles

Figure 20 contains a time plot of the changes in cluster assignments over time as the trial progresses. We see certain clusters, such as four, nine, and 13, where the participant's eyes spend a relatively longer amount of time in at the end of the cluster. These correspond to behaviors where the eyes are fixating around the targets. For most of the other clusters, Participant B's eyes only spend a short duration at a time in the cluster and the clusters are revisited multiple times. This indicates that the clusterings are not very distinct.

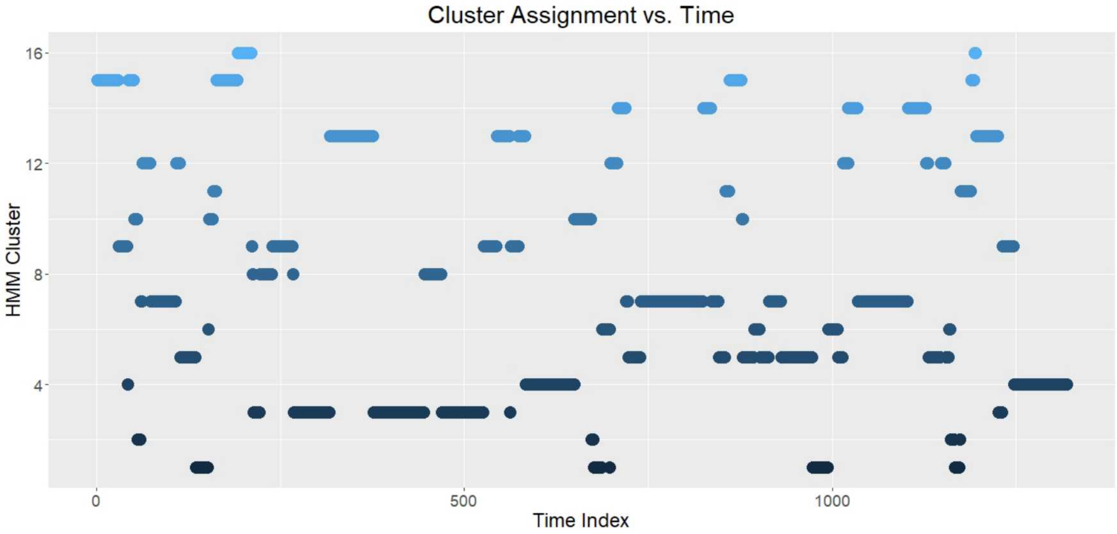


Figure 20. Time Plot of Cluster Assignments for Participant B with covariates lenratio, angle, anglediffs, and totalAngles

4.4.3 HMM Clustering Results with Covariate angle

We fit a HMM to the eye movement data with angle as the lone covariate. Using the Bayesian information criterion (BIC), we find the optimal HMM is one with 16 hidden states or clusters. We will now analyze the five aforementioned visualizations for Participant B's data.

Figure 21 contains the cluster assignment plots (without (a) and with (b) the 95% confidence ellipses for each cluster). In these plots, the four targets are indicated by the larger colored percent signs. Similar to what is seen with Participant A, the ellipses seen in Figure 21(b) indicate that there are much fewer overlapping clusters as compared to when multiple covariates are incorporated. This indicates that distinct patterns in the eye movement data can be identified better by the HMM. However, the clusters are not quite as distinct as when no covariates are included. We note that unlike the models with no covariates and with multiple covariates, where we observe a pretty horizontal clustering along the bottom part of the image (around y=83 and spanning $x \in (0, 75)$) and across the right ride of the image (spanning $x \in (75, 100)$ and $y \in (20, 100)$), we do not see the same in the model with only the angle covariate. With the angle covariate model, we see a cluster across the bottom of the image that has more of an oblique positive slope and doesn't span as far horizontally (corresponds to cluster 10), spanning $x \in (20, 63)$ and $y \in (65, 90)$. Across the right side of the image, there is one clear clustering (corresponds to cluster 11), but it is wider than the observed clusterings in the other two models.

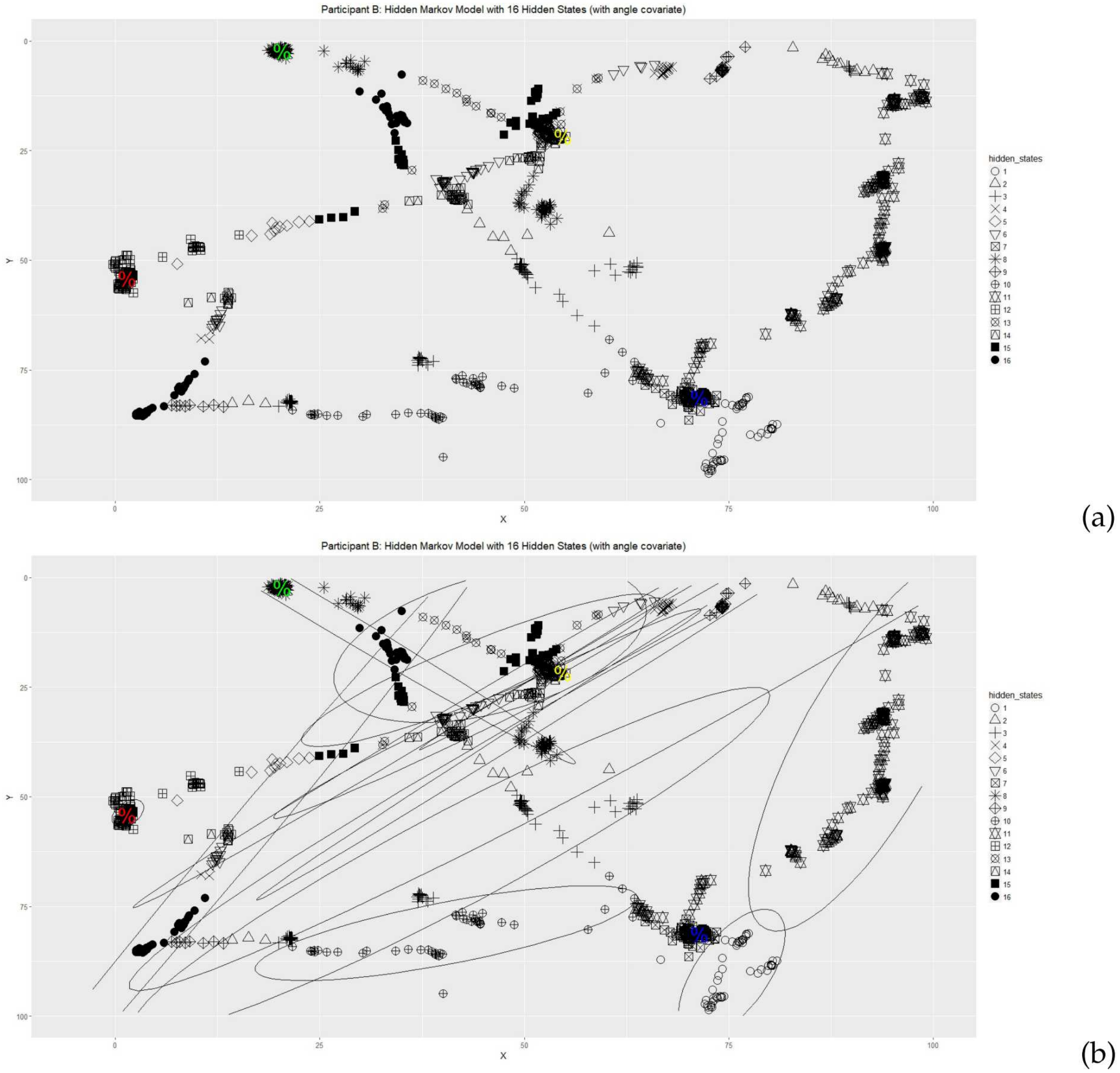


Figure 21. Cluster assignment plot(a) and cluster assignment plot with 95% confidence ellipses for each cluster (b) for Participant B with covariate angle

Figure 22 contains the uncertainty plots (without (a) and with (b) the 95% confidence ellipses for each cluster). We see more clearly from Figure 22 (b) that most of the points have been clustered to their assigned clusters with relatively low uncertainty, as indicated by the fact that most data points are marked with dark blue colors. There is a small of points in the lighter blue colors that have higher classification uncertainty, but there does not appear to be a systemic group of points that have higher uncertainty. While there is still some overlap between the ellipses, the clusters are much more distinct when angle is the only covariate considered, as opposed to when many more covariates are considered.

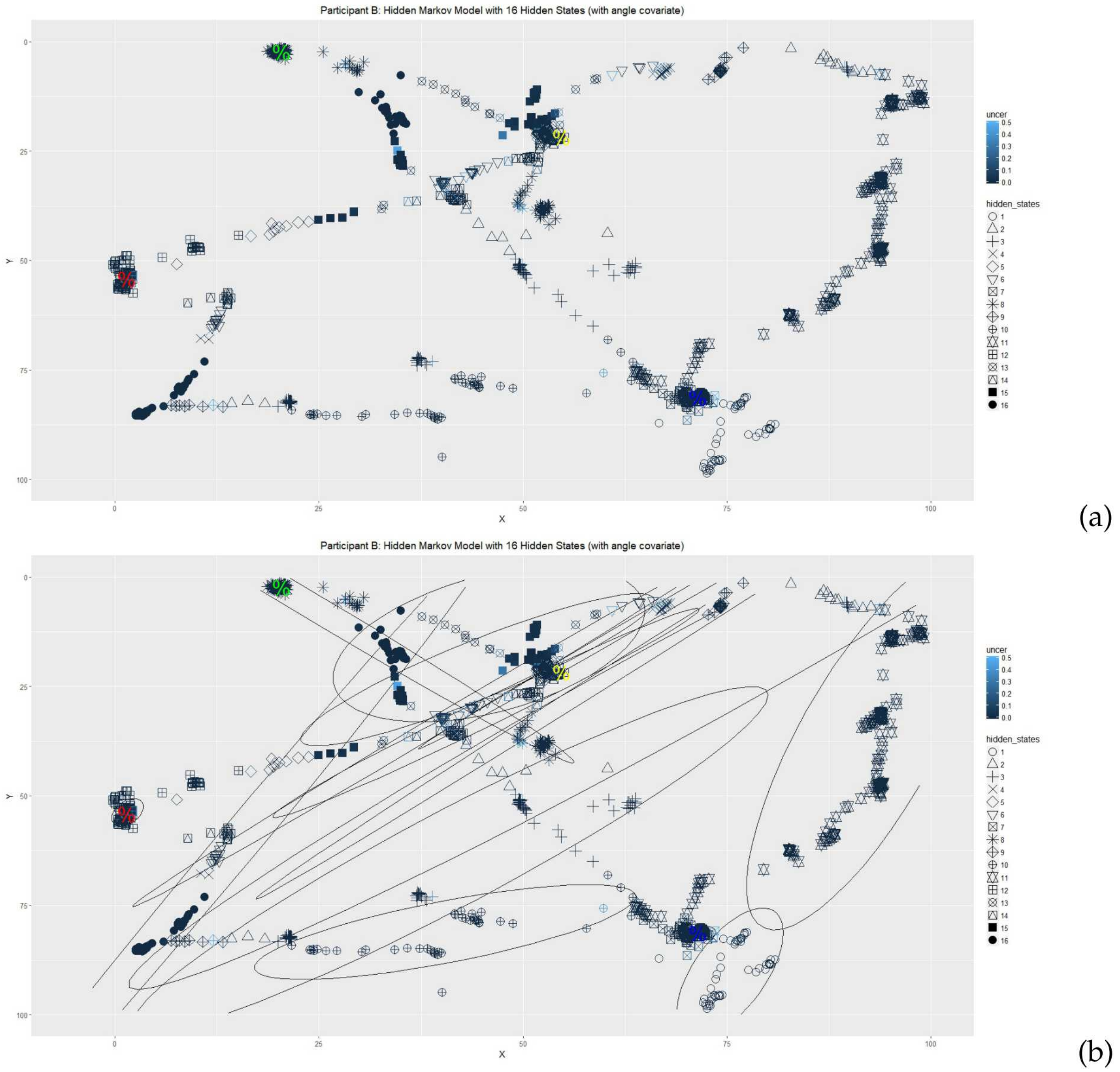


Figure 22. Uncertainty plot(a) and uncertainty plot with 95% confidence ellipses for each cluster (b) for Participant B with covariate angle

Figure 23 contains a time plot of the changes in cluster assignments over time as the trial progresses. We see certain clusters, such as seven, eight, 11, and 12, where the participant's eyes spend a relatively longer amount of time. These correspond to when the subject is fixating on the target or in a continuous movement from one location in the image to another. There are also groups of clusters, such as clusters 13-16, that seem to be visited all within a certain time range and their visit times are complementary to one another. This indicates that these clusters appear to describe eye movement behavior in a similar area of the image.

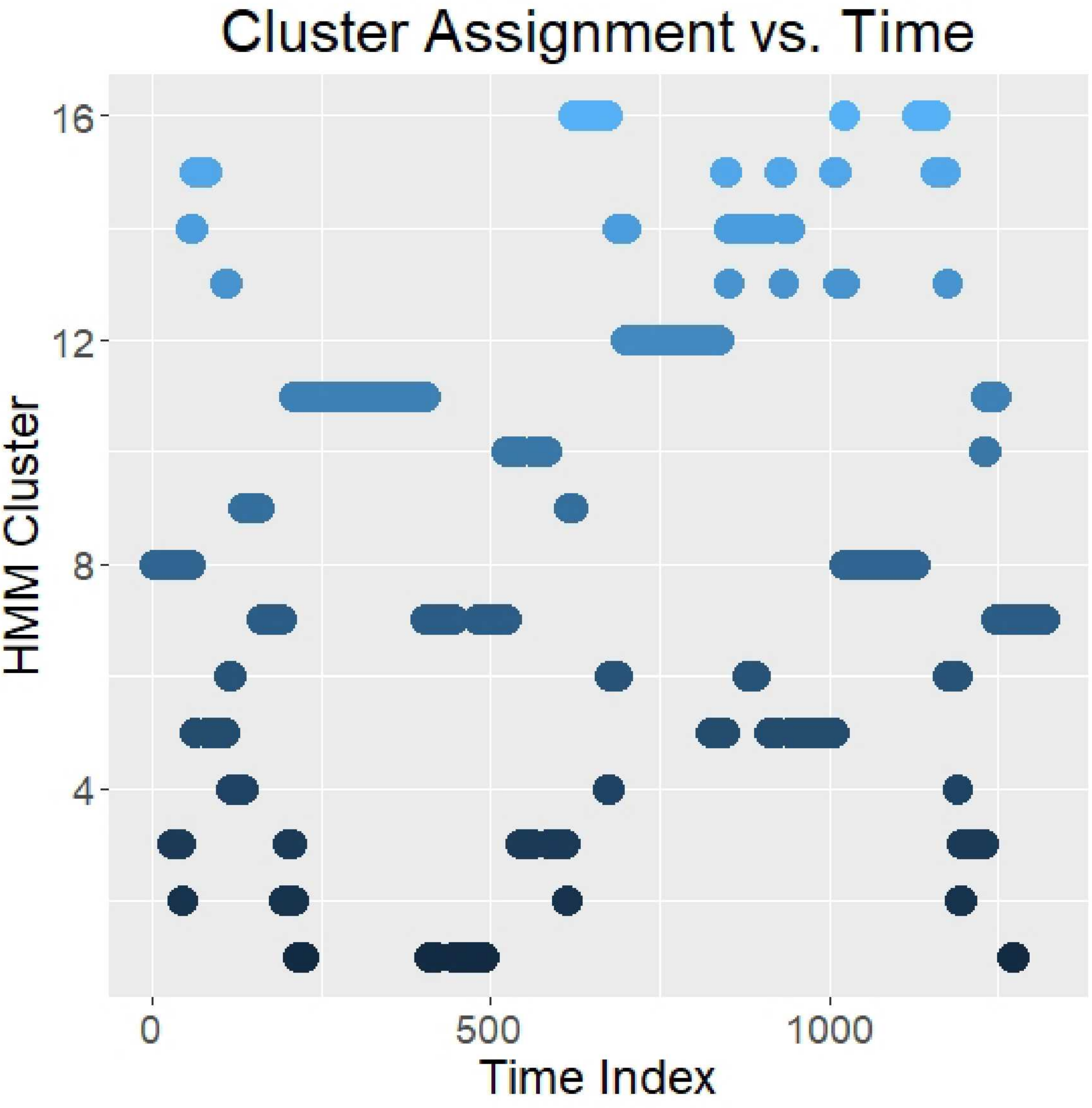


Figure 23. Time Plot of Cluster Assignments for Participant B with covariate angle

4.4.4 Pairwise Comparisons of Clusterings

The Dunn index values in Table 3 reveal the model with the angle covariate has the highest separability, followed by the model with no covariates. The Dunn index values confirms our findings in the visualizations that the model with multiple covariates had much more overlap between clusters. However, from the visualizations, it is not clear that the model with the angle covariate has more separability between clusters than the model with no covariates. The visualizations reveal the specific nature of the separability between clusters (i.e. which clusters are more distinct than each other), which would not be revealed in the Dunn index.

Covariates	Dunn index
None	0.001408568
Angle	0.00156004
Multiple	7.012993×10^{-5}

Table 3. Dunn Index Values Measuring Separability of Clusters for Participant B's HMM Modelings

The results for the external evaluation measures in Table 4 indicate that the models with no covariates and no covariates have the most similarity. From visual inspection, that would not appear to be the right conclusion, as the clusterings for multiple covariates and angle covariate would appear to be a bit more similar, with both models resulting in multiple clusters with upward sloping ellipses. However, it would appear these results are trumped by the similarities in the clusters across the bottom part and right side of the image in the models with no covariates and multiple covariates.

Table 4. Pairwise Comparisons of Similarities Between Participant B's HMM Modelings.

Model 1 Covariates	Model 2 Covariates	Rand	HA	MA	FM	Jaccard
None	Angle	0.8342803	0.2559540	0.2597948	0.3549258	0.2082241
None	Multiple	0.8561873	0.3417296	0.3452172	0.4330626	0.2634373
Angle	Multiple	0.8738549	0.2536251	0.2590912	0.3232725	0.1926724

4.5 Comparison of Participants A and B

In addition to discovering patterns in time-series data, such as the eye movement behaviors of persons one and two, we can use the visualizations to compare the clusterings of the two datasets for Participants A and B. We compare the clusterings of persons one and two's eye movement patterns when no covariates are included in the respective HMMs. For the sake of convenience, we reproduce the plots of the eye movement data for Participants A and B, cluster assignments with ellipses, uncertainty with ellipses, and time plots for persons one and two, and analyze the similarities and differences between the two persons' eye movement behaviors.

Comparing the plots of the eye movement data for Participants A and B (Figure 24), we see that Participant A has a cleaner eye movement pattern and takes straighter paths to each target, while Participant B has some more circuitous paths and the eye movement pattern covers more area of the image.

The cluster assignment plots for the two participants (Figure 25) confirm the findings from the plots of their eye movement data. We see that Participant A has more distinct clusterings, with generally smaller clusters, than Participant B.

In the uncertainty plots (Figure 26), we see that the data points with higher classification uncertainties are the points where the participants tends to move over the same area more than once. It is plausible that these are the areas the persons had the most difficulties in locating the targets and deciding how to proceed. For Participant A, this is most evident in the top left hand corner of the image. For Participant B, this is most evident in the middle of the image.

The time plots for the two participants (Figure 27) confirm our findings with the three previous set of plots that both Participants A and B have generally clean eye movement patterns in finding the four targets. We draw this conclusion because we do not see many clusters being revisited by either participant. However, we can conclude that Participant A has a cleaner eye movement pattern overall because there are fewer clusters being revisited, and we can isolate longer periods of time being spent in one cluster and distinguishing the eye movement behavior represented by the cluster. For Participant A, we can distinguish that cluster two is the participant's eyes moving in and fixating on target four (represented by the blue percent sign), while clusters three, four, and nine all represent segments of the participant's eyes moving towards and fixating on target one (represented by the red percent sign). For Participant B, we can distinguish cluster four as the behavior of the participant's eyes moving across the bottom of the image from left to right towards target four and then fixating on the target. We also distinguish clusters six, 12, and 10 as helping describe the participant's eye movement from target one to target two (represented by the green percent sign).

We note that we cannot use the external evaluation measures for clusterings to compare the clusterings of Participants A and B because those measures compare two clusterings of the same dataset, and therefore, require that the two datasets be of the same length. Since that is not the case for Participants A and B, we cannot use those measures to supplement the analysis done with the visualization tools.

5 DISCUSSION

We introduce five different tools for visualizing the clustering results and the classification uncertainties at each data point of a dataset consisting of multivariate observations that are correlated in time. We demonstrate our visualization tools for clustering an eye tracking dataset, which tracks the spatial location of where a participant's eyes are looking at on an image while searching for pre-defined targets on the image, into segments of patterns of eye movement behavior over the course of a task where a participant is looking for four targets. While we apply our visualization tools to the task of clustering eye tracking data with the HMM, these tools can be applied to the task of isolating behaviors in any multivariate time-series dataset using any kind of probabilistic clustering model that is suitable for time-dependent data.

We plot not only the locations of the eye movement data, but we look at the clusters each data point is classified into (the cluster assignment plot), the classification uncertainty for each data point (uncertainty plot), and a time plot which tracks the cluster that data points are assigned to over time. We also add the 95% confidence ellipses for each cluster to the cluster assignment and uncertainty plots. Adding the ellipses to the cluster assignment plot allows the visualization of the sizes of each cluster to see how tightly bound the clusters are and overlap between clusters. Adding the ellipses to the uncertainty plot allows the visualization of the confidence level of the classifications and if the classification uncertainties are related to the data points position relative to the centroids of the clusters. The time plot allows us to see how the cluster assignments change over time, in particular how much time are spent in each cluster and if clusters are revisited. Combined with the cluster assignment and uncertainty plots, we can infer patterns in the data (such as eye movement patterns in different parts of the image and if the participant has found and the eyes are fixating on a target) and gauge the quality of the clustering done by the HMM.

In addition to finding patterns in time-series datasets, the visualizations can also be used to compare clusterings when covariates are incorporated into the model and clusterings with different (but related datasets). We demonstrate these comparisons by incorporating features that help describe the trajectory of eye movement data into the HMM and see how they can affect the quality of the clustering. We also compare the clustering results for two different participant that complete the same eye tracking task.

We note that our analysis done through our visualization tools agree with the cluster evaluation measures that we compute. Our conclusions for the separability of clusterings done by a HMM agree with the Dunn index we compute for each HMM we fit. Our conclusions for the pairwise comparison of two clusterings agree with the values of the five external evaluation measures (Rand index, Hubert and Arabie's adjusted Rand index, Morey and Agresti's adjusted Rand index, Fowlkes and Mallows' index, and Jaccard index) we compute. We note that all of these quantitative measures are for an entire dataset. The Dunn index measures the separability between all clusters fitted by a HMM over an entire dataset, but it doesn't pinpoint the exact separabilities between any specific clusters. Our visualizations allow us to determine specifics of distinctions between clusters and the sizes of individual clusters. The five external measures compare two clustering models as a whole, but they don't specify individual similarities and differences between the two clustering models' results. Our visualizations allow us to make these inferences, which illustrates the ability of the visualization to improve the information exploitation results of our datasets.

6 ANTICIPATED OUTCOMES AND IMPACTS

The main value of this work is the marked value gained by including visualizations of the clustering models in addition to the standard, global numerical measures. Having that additional evaluation method allows for: (1) faster, more intuitive ability to see how well a model is fitting (e.g., visualization for the models with all the covariates) and (2) additional information as to why the model is or is not fitting appropriately (e.g., overlap of 95% confidence ellipses, revisits through time, distribution of uncertainty of data point assignment).

While the models do not fit our given eye tracking dataset well, having the visualizations allow us to more quickly determine that and gain insight into why that is the case. A clear future step is to apply these techniques to a data set in which the HMM methods used here led to better clusters, and see if we also find utility in including the visualizations. Another related future step is to determine what kind of models best fit our given eye tracking dataset.

We have identified several areas of open methodology problems related cluster analysis of multivariate time-series data. First, there are currently no quality goodness-of-fit statistics for mixture models. While there are good tests for normality, such as the Kolmogorov-Smirnov and Anderson-Darling tests, these do not assess whether or not the clustering of the data points is good. They would only be able to assess whether or not the clusters follow a normal distribution. The second is integrating cluster separability measures, such as the Dunn index or the measure proposed by (30), into the computation of classification uncertainty, which is heavily based on the classification posterior probability estimated via the Baum-Welch algorithm for HMMs or the EM algorithm for other probabilistic clustering models. There is empirical evidence that lower separability between clusters can lead to higher classification uncertainties for data points. The third is extending current measures for classification uncertainty at individual data points to quantifying the uncertainty of clusterings, and then visualizing these uncertainty bounds. This will greatly enhance our ability to evaluate whether or not a model is doing a good job of clustering the data. Fourth, as we have noted in section 4.1, we only focus on one task out of a larger eye tracking dataset that contains eye movement data asks participants to complete four tasks. The entire dataset from this study consists of 16 participants completing four tasks and doing five trials per task. The question of how to incorporate multiple tasks and trials, as well as multiple participants, in order to draw conclusions about differences between participants or populations of participants remains an open problem. This open problem can be generalized to integrating and fully exploiting the information from multiple related time-series datasets. There are other application areas, such as finance and biomedical imaging, that contain these kinds of datasets. Fifth, we need to develop useful visualizations when clustering time-series data with datapoints of more than two dimensions. In this article, we conveniently analyze spatial data, which has two coordinates, and we can plot the data and their clusterings directly. While most visualizations of multivariate data points of more than two dimensions plot the first two principal components, it is unclear whether or not that is the most effective visualization for every application. Finally, we need to extend the current measures for classification uncertainty at the level of individual data points to quantifying the uncertainty of clusterings, and then visualizing these uncertainty bounds.

Another direction relevant to this work is to further dig into the geospatial temporal domain. The HMM methods used here do not work well for our eye tracking data set on SAR imagery analysts. The questions of what would be a more useful approach remains. The current methods force every data point into a cluster; however, traditional eye tracking techniques drop data that do not align to meaningful eye movement patterns. That approach makes a lot of sense, since every sample at 60 Hz taken from a (somewhat noisy) eye tracking machine is not always a data point of value. Sometimes ones eyes are in the middle of a large movement and that point might be relatively arbitrary; or perhaps that point is captured when a participant glances away from the screen; or perhaps the eye tracker drifts for a moment. Being able to identify and discard those points would allow for cleaner, stronger patterns to be pulled from the data.

The techniques used in this project are also intentionally entirely bottom-up, data-driven, unsupervised methods. We find that the clusters that resulted do not tie to meaningful eye movement patterns and therefore are not particularly useful in addressing our question of interest for that dataset: what are the decision-making patterns of interest in SAR imagery analysts? While we intentionally focus on that bottom-up approach to allow the data to speak for itself and not need to hand-code the eye movement data, we believe it might be helpful to impose a structure that allows for guidance from top-down components of meaningful eye movement patterns (e.g., fixation), while allowing that structure to be filled in from a bottom-up perspective. Future work could dig into this complex need, and it would be quite helpful in many geospatial temporal data mission domains here at Sandia, such as the decision-making behavior of SAR imagery analysts.

Another element we would like to further address is to better pull in the temporal information in addition to the spatial information. The HMM methods used in this project were biased toward accounting for the spatial information more than the temporal informationit only looked one point back to pull in temporal information. For a 60 Hz eye tracker, thats only temporally accounting for information 16.7 ms back. However, seeing and reacting to information you see can take around 250 ms. Being able to better incorporate temporal information for geospatial temporal data like eye movement patterns would allow us to better take advantage of the full spectrum of meaningful information in that data. Related to better bringing in the temporal information is determining how spatial and temporal sources of information contribute to the analysis results, and decomposing overall error and attributing error to these various sources.

One final area of future work is related to semi-supervised probabilistic clustering of time-dependent data, which clusters data when limited to no training data or covariates are available and patterns between and within vectors of time-dependent data can be isolated. A probabilistic clustering method that clusters dependent data and factors in partial covariate or training information will need to be developed. Afterwards, methods to visualize clustering based on the covariates in order to gauge the influence of covariates will need to be developed. This is analogous to visualizing a regression model, where the relationship between an independent and dependent variable can be visualized and determined. This work can make an impact in many Sandia mission areas that contain time-dependent data and have a need to determine influences of other data sources that may be incomplete. Examples of mission areas where developed methods can be applied to include network traffic data to determine indicators of malicious behavior, as well as wearable device data collected from the DTRA funded Sandia-led WATCH project to determine physical health indicators of Grand Canyon hikers.

We have identified many possible research directions related to probabilistic clustering methods, uncertainty quantification, and visualization methods. We have also identified several mission areas that can benefit from these methods, such as visual information foraging for eye tracking data, network data, and time-dependent data collected from wearable devices. These research directions can be the foundation for a long-term partnership with NGA-R because it spans many of their interest areas, such as sensors, geospatial and cyber data, anticipatory analytics, environment and culture, space, and automated analytics. Our methods can be used on many of their primary datasets, such as geo-spatial imagery, and can be used to visualize new types of datasets they are interested in. In addition, our interest in determining error sources (such as spatial and temporal information) and pulling in semantic meaning into probabilistic clustering methods fits into NGA-R's interest in understandable and comprehensible machine learning methods. This fits into an overall collaboration plan that spans fundamental research to deployable capabilities.

7 CONCLUSION

We introduce five different tools for visualizing the clustering results and the classification uncertainties at each data point of a dataset consisting of multivariate observations that are correlated in time. The additional value of the visualizations that are not provided by the standard quantitative global measures are highlighted in two main areas: (1) faster, more intuitive ability to see how well a model is fitting (e.g., visualization for the models with all the covariates) and (2) additional information as to why the model is or is not fitting appropriately (e.g., overlap of 95% confidence ellipses, revisits through time, distribution of uncertainty of data point assignment). In addition to the separability between clusters within a clustering, as measured by internal measures such as the Dunn index, and the difference between the performance of two clustering models, as measured by external measures such as the adjusted Rand index, the visualizations provide more specific information that allows determination of separation between specific clusters or visualizing the specific differences between two clustering models. The visualizations provide not only provide information on the performance of probabilistic clustering models, but it also allows for the comparison of the clusterings of data from multiple subjects or data sources.

We have identified many areas of future work that can be built upon the work presented in this report. A number of methodology problems yet to be addressed include determining the goodness-of-fit of probabilistic clustering models, incorporating separability into clustering algorithms, and expanding visualization tools to visualize clustering results for data of more than two dimensions and to compute and visualize the uncertainty of clusters. We also need to figure out how to identify clustering and uncertainty associated with spatial and temporal characteristics of data, as well as incorporating subject-specific information so that probabilistic clustering models, which are data-driven and unsupervised, can provide more interpretable results.

We have identified multiple Sandia mission areas and other application areas where data analysis can be enhanced by these visualizations because time-dependent data is central in these areas. In all of these areas, the ability to detect trends both within and between series of time-dependent data, as well as the measure of the level of uncertainty of the clustering results, not only improves data exploitation capabilities, but the uncertainty information also provides decision-makers with a measured level of confidence that they should have in the analysis presented to them.

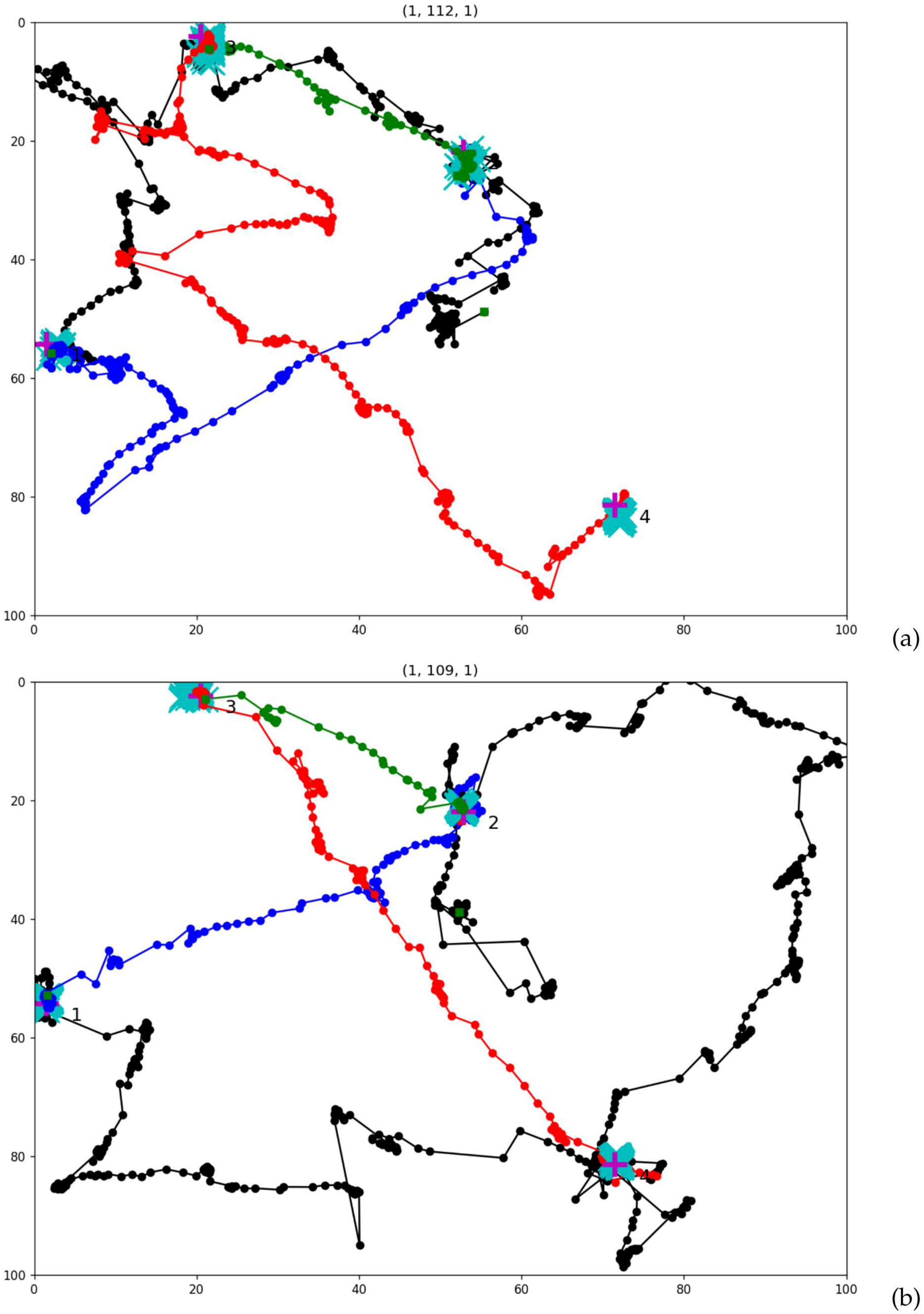


Figure 24. Plots of Eye Movement Data for Participants A (a) and B (b)

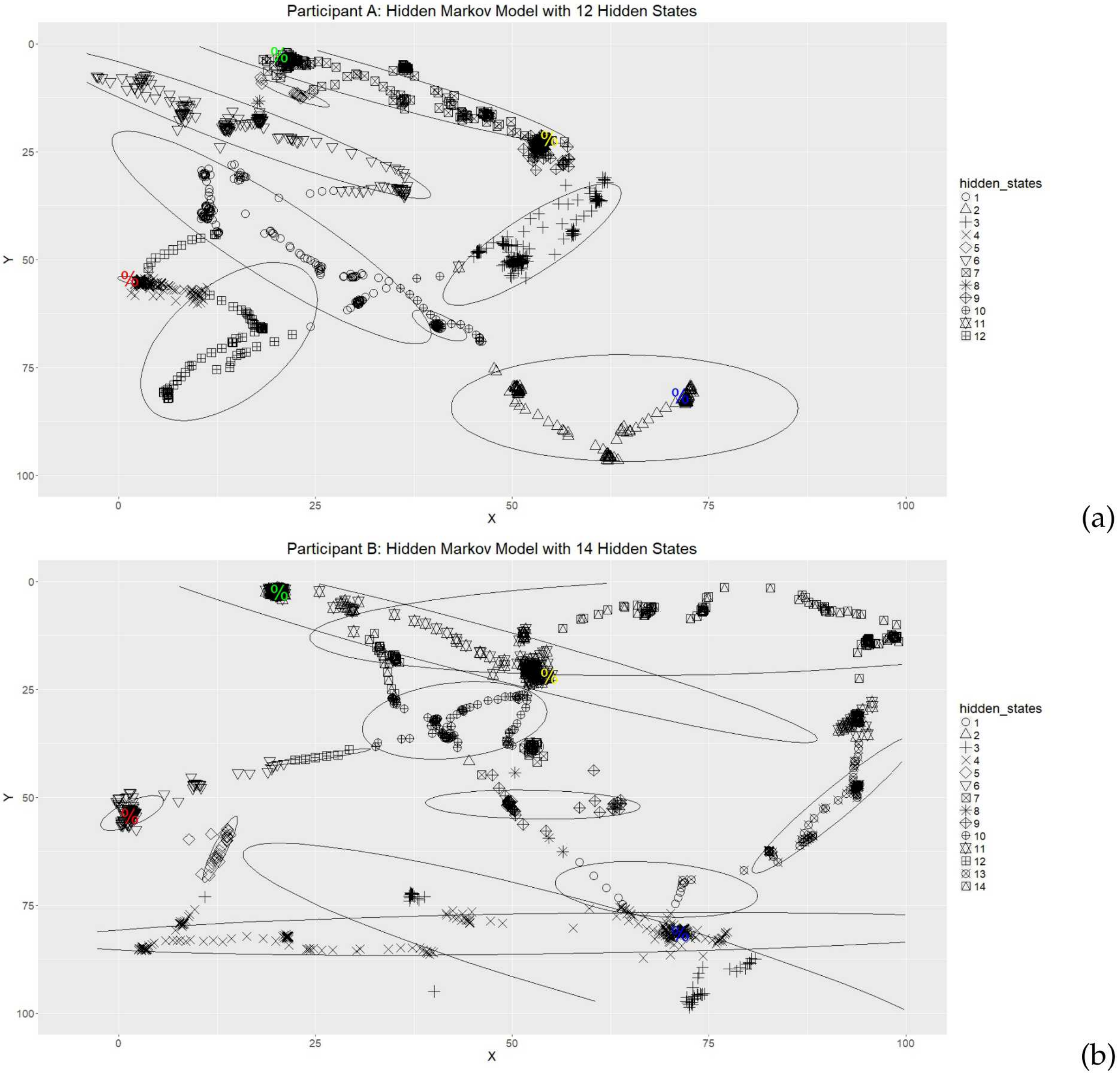


Figure 25. Cluster Assignment Plots with 95% Confidence Ellipses for Participants A (a) and B (b)

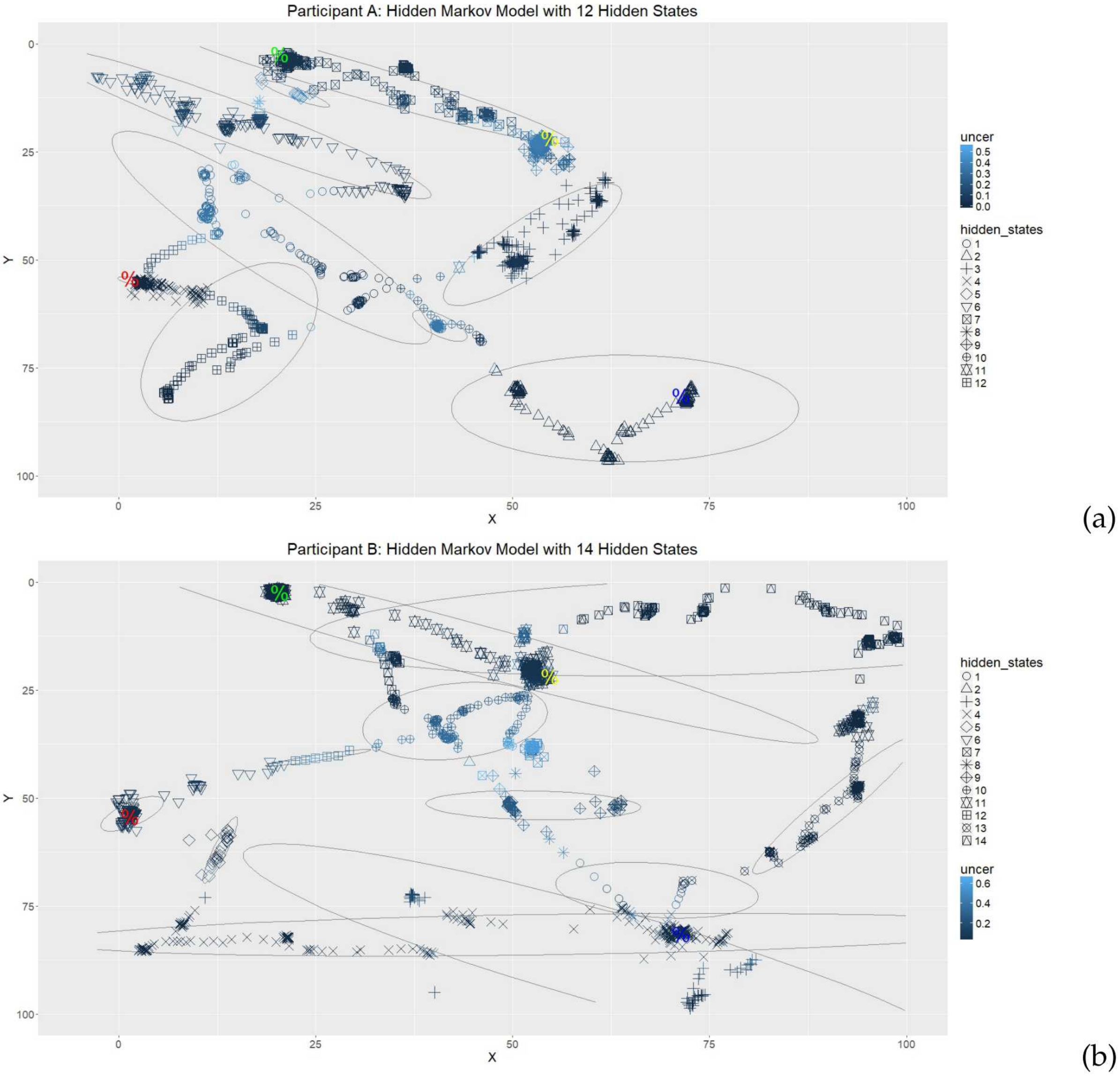


Figure 26. Classification Uncertainty Plots with 95% Confidence Ellipses for Participants A (a) and B (b)

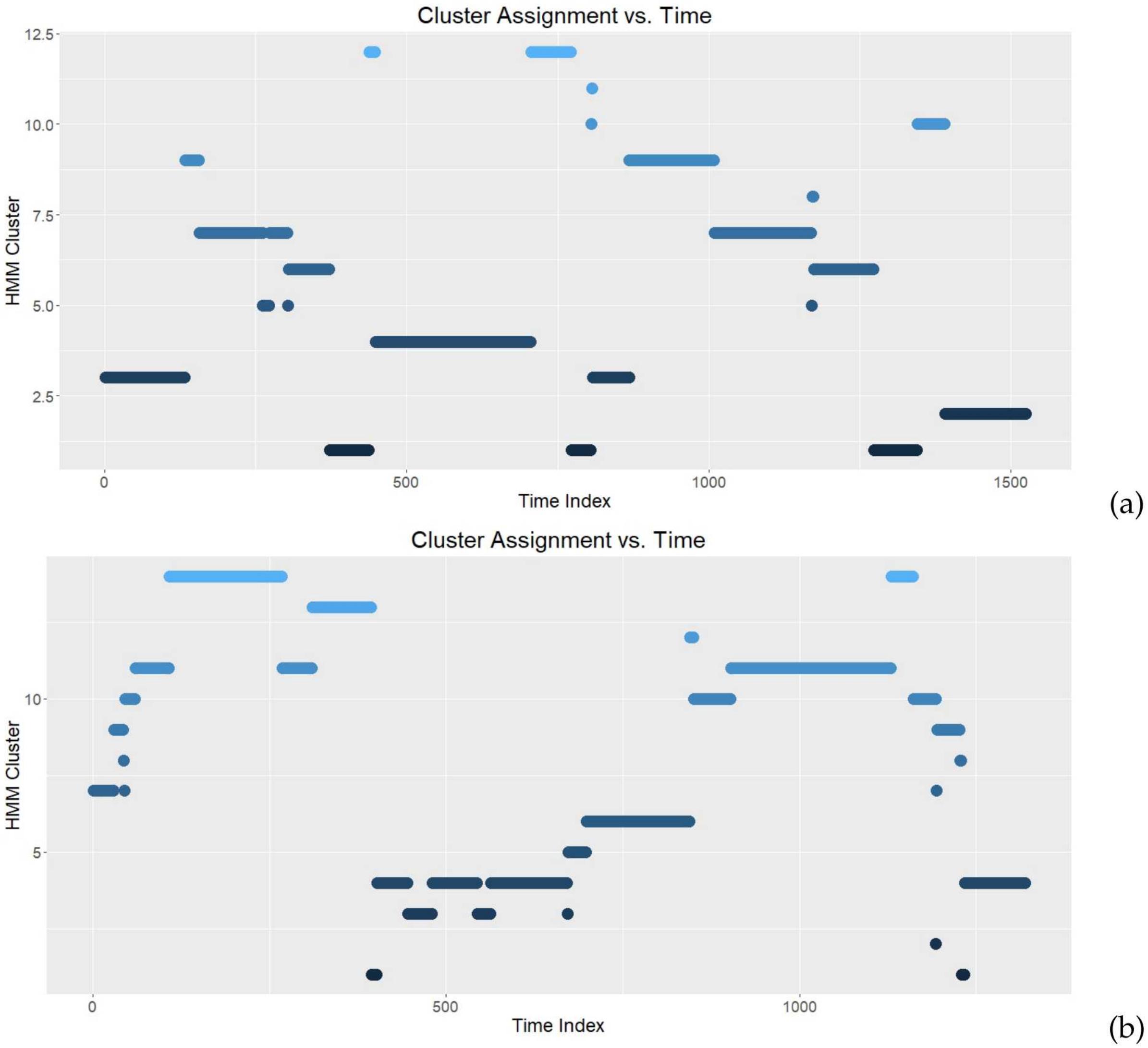


Figure 27. Time Plots for Participants A (a) and B (b)

References

[1] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 49–60, New York, NY, USA, 1999. ACM.

[2] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[3] Leonard E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In Oved Shisha, editor, *Inequalities III: Proceedings of the Third Symposium on Inequalities*, pages 1–8, University of California, Los Angeles, 1972. Academic Press.

[4] Tatiana Benaglia, Didier Chauveau, and David Hunter. Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger, World Scientific Publishing Co.*, pages 15–27, 2011. ¡hal-00353297¿.

[5] Tatiana Benaglia, Didier Chauveau, and David R. Hunter. An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[6] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, 2002.

[7] S. de Haan-Rietdijk, P. Kuppens, C. S. Bergeman, L. B. Sheeber, N. B. Allen, and E. L. Hamaker. On the use of mixed markov models for intensive longitudinal data. *Multivariate Behavioral Research*, 52(6):747–767, 2017. PMID: 28956618.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[9] Krisin M. Divis, Maximillian G. Chen, Laura A. McNamara, J. Dan Morrow, and David N. Perkins. Moving from low level eye movement data to meaningful content in dynamic environments. 2017 European Conference on Eye Movements Poster, August 2017.

[10] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters

in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[12] LAWRENCE R. RABINER FELLOW and IEEE. A tutorial on hidden markov models and selected applications in speech recognition. In Alex Waibel, , and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 267 – 296. Morgan Kaufmann, San Francisco, 1990.

[13] Dennis Fisher and Paul Hoffman. The adjusted rand statistic: A sas macro. 53:417–423, 02 1988.

[14] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[15] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[16] Chris Fraley and Adrian E. Raftery. Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, 18(6):1–13, 1 2007.

[17] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2):245–273, Nov 1997.

[18] K. Ghidouche, T. Kechadi, and A. K. Tari. Study and analysis of temporal data using hidden markov models. In *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, pages 16–20, June 2011.

[19] Wenbing Gong, Shenghui Fang, Guang Yang, and Mengyu Ge. Using a hidden markov model for improving the spatial-temporal consistency of time series land cover classification. *ISPRS International Journal of Geo-Information*, 6(10):292, Sep 2017.

[20] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. *SIGMOD Rec.*, 27(2):73–84, June 1998.

[21] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[22] P. Jaccard. *Nouvelles recherches sur la distribution florale*. Bulletin de la Société vaudoise des sciences naturelles. Impr. Réunies, 1908.

[23] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Proessing, Computational Lingistics, and Speech Recognition*, chapter 9: Hidden Markov Models, pages 122–141. Third edition draft edition, 2017.

[24] Chang-Jin Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

[25] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480, 1990.

[26] Anders Krogh. An introduction to hidden markov models for biological sequences. 32, 12 1998.

[27] ROLF LANGEHEINE and Frank van de Pol. A unifying framework for markov modeling in discrete space and discrete time. 18:416–441, 05 1990.

[28] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006.

[29] Theodore C Lystig and James P Hughes. Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689, 2002.

[30] Ranjan Maitra and Volodymyr Melnykov. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, 19(2):354–376, 2010.

[31] Jean-Francois Mari, Florence Le Ber, El-Ghali Lazrak, Marc Benoit, Catherine Eng amd Annabelle Thibessard, and Pierre Leblond. Using markov models to mine temporal and spatial data. *New Fundamental Technologies in Data Mining, Intech*, (978-953-307-547-1):561584, 2011.

[32] Jean-Franois Mari and Florence Le Ber. Temporal and spatial data mining with second-order hidden markov models. *Soft Comput.*, 10:406–414, 2006.

[33] Paul McNicholas and Thomas Murphy. Model-based clustering of longitudinal data. 38:153 – 168, 03 2010.

[34] Volodymyr Melnykov. On the distribution of posterior probabilities in finite mixture models with application in clustering. *Journal of Multivariate Analysis*, 122:175 – 189, 2013.

[35] George A. Miller. Finite markov processes in psychology. *Psychometrika*, 17(2):149–167, Jun 1952.

[36] George A. Miller and Noam Chomsky. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology*, volume II, pages 419–491. John Wiley, New York, 1963.

[37] Leslie C. Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.

[38] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[39] Erich Schikuta. Grid-clustering: A fast hierarchical clustering method for very large data sets. Technical report, 1993.

[40] Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.

[41] Anthony K. H. Tung, Raymond T. Ng, Laks V. S. Lakshmanan, and Jiawei Han. Constraint-based clustering in large databases. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, pages 405–419, Berlin, Heidelberg, 2001. Springer-Verlag.

[42] Ingmar Visser, Maartje E. J. Raijmakers, and Han L. J. van der Maas. *Hidden Markov Models for Individual Time Series*, pages 269–289. Springer US, New York, NY, 2009.

[43] Ingmar Visser and Maarten Speekenbrink. depmixs4: An R package for hidden markov models. *Journal of Statistical Software, Articles*, 36(7):1–21, 2010.

[44] Silke Wagner and Dorothea Wagner. Comparing clusterings- an overview, 2007.

[45] Ying Wang, Susan M. Resnick, and Christos Davatzikos. Spatio-temporal analysis of brain MRI images using hidden markov models. In *MICCAI (2)*, volume 6362 of *Lecture Notes in Computer Science*, pages 160–168. Springer, 2010.

[46] Thomas D Wickens. *Models for behavior : stochastic processes in psychology*. San Francisco : W.H. Freeman, 1982. Includes index.

[47] Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, Anaid Flesken, and Guido Sanguinetti. *Modeling Conflict Dynamics with Spatio-temporal Data*. Springer International Publishing, 1 edition, 2013.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)

Sandia National Laboratories