This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Sandia National Laboratories

SAND2019-8208C

Tallis: Dimension Reduction for Mixed-Type Data

Robin Tu (Org. 09365), rtu@sandia.gov
Alexander Foss (Org. 09136), afoss@sandia.gov

Introduction:

Dimensionality reduction involves constructing a parsimonious set of variables that explain a large multivariate data set. Current dimensionality reduction techniques mostly focus on continuous data types and do not allow for mixed-type data beyond elliptically symmetric family of distributions (Normal, t, Laplace, etc.).

Tallis Goals:

- *Dimension Reduction of mixed-type data*: Ingesting multiple heterogeneous data streams simultaneously– Variables are allowed to be skewed, count, categorical, multi-outcome data (multinomial and simplex).
- *Anomaly Detection*: Build a flexible *interpretable* model for "normal" behavior and flag "abnormal" behavior with some form of uncertainty quantification.
- *Discriminant Analysis*: Build a separate model for each of *k* classes, and assign observations to the most likely class.

Method:

- Factor model with *k* latent components
- Model is fit with the EM algorithm
  - E step: Estimate latent factors via Monte Carlo integration.
  - M step: Estimate matrix of factor loadings using nonlinear non-normal regression.

$$\underset{q_1\times 1}{\vec{X}_{t,1}} = \vec{\mu}_1(\underset{d_1\times k}{\Lambda_1}\overset{k\times 1}{\vec{Z}_t}) + \underset{q_1\times 1}{\vec{U}_{t,1}}$$

$$\vdots$$

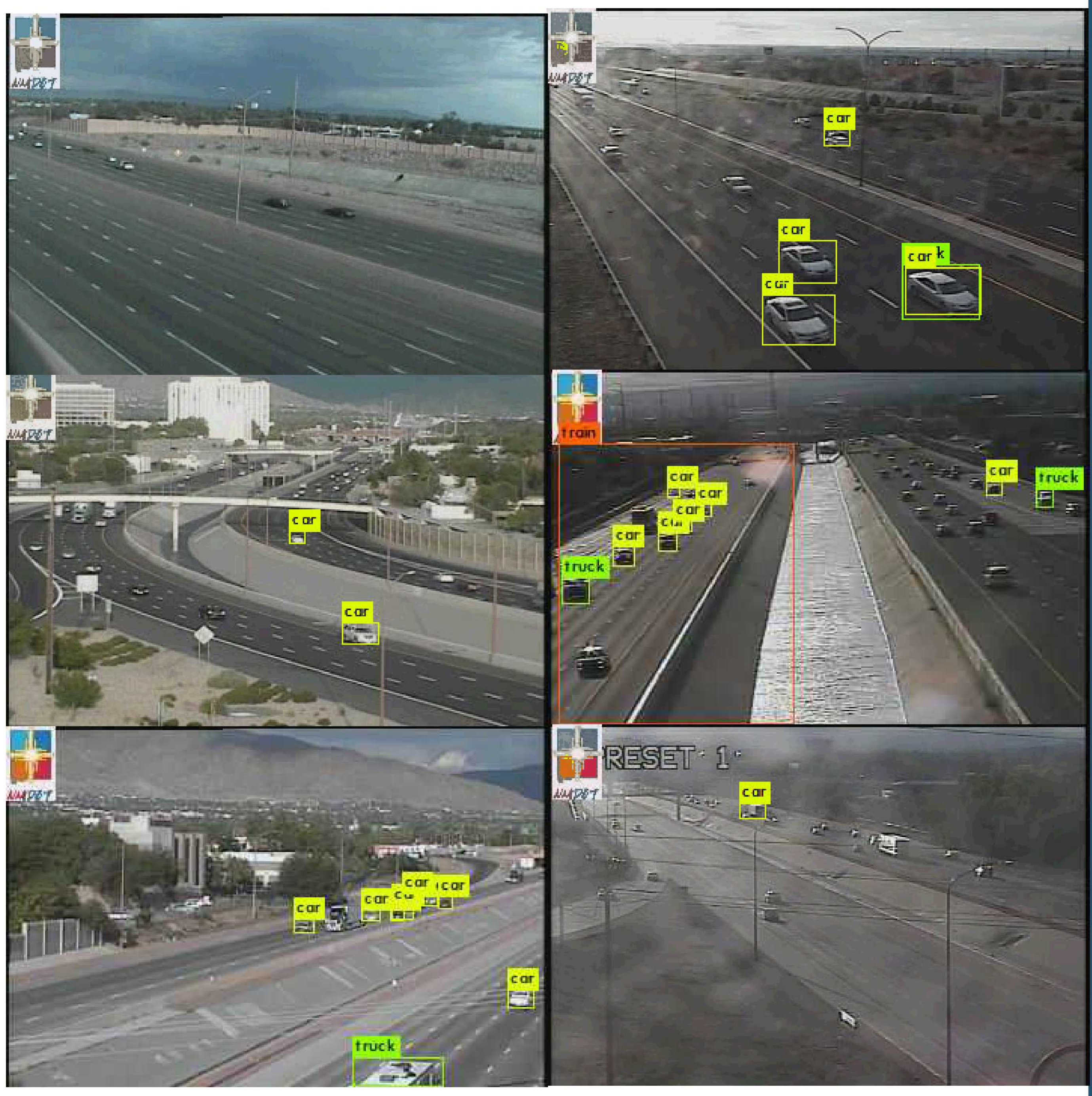
$$\underset{q_p\times 1}{\vec{X}_{t,p}} = \vec{\mu}_p(\underset{d_p\times k}{\Lambda_p}\overset{k\times 1}{\vec{Z}_t}) + \underset{q_p\times 1}{\vec{U}_{t,p}}$$

$$\vec{Z}_t \sim MVN(\vec{\nu}(t), I_k)$$

- k: Number of latent factors, $k < \sum_i q_i$
- d_i: Number of parameters estimated for variable i
- $\vec{\mu}_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{q_i}$: mean function for the i^{th} input vector
- Λ_i: matrix of factor loadings
- $\vec{Z}_t$: time-dependent vector of latent factors
- $\vec{U}_{t,i}$: vector of residuals
- I_k: $k \times k$ identity matrix
- $\vec{\nu}(t)$: time-dependent mean vector

(Below) Weather data

Time	Temp	Dew Pt.	Humidity	Wind Dir	Wind Spd	Wind Gust	Pressure	Precip.	Precip Accum	Condition
6:52 PM	80 F	51 F	36%	S	13 mph	0 mph	24.8 in	0.0 in	0.0 in	Cloudy
7:52 PM	79 F	50 F	36%	S	10 mph	0 mph	24.8 in	0.0 in	0.0 in	Cloudy



(Above) ML algorithm applied to NMDOT camera's of the I-40, 7/14/19 at 7:20pm (top left) at University, (top right) at Carlisle, (middle left) at San Mateo, (middle right) at Louisiana, (bottom left) at Pennsylvania (bottom right) at Eubank

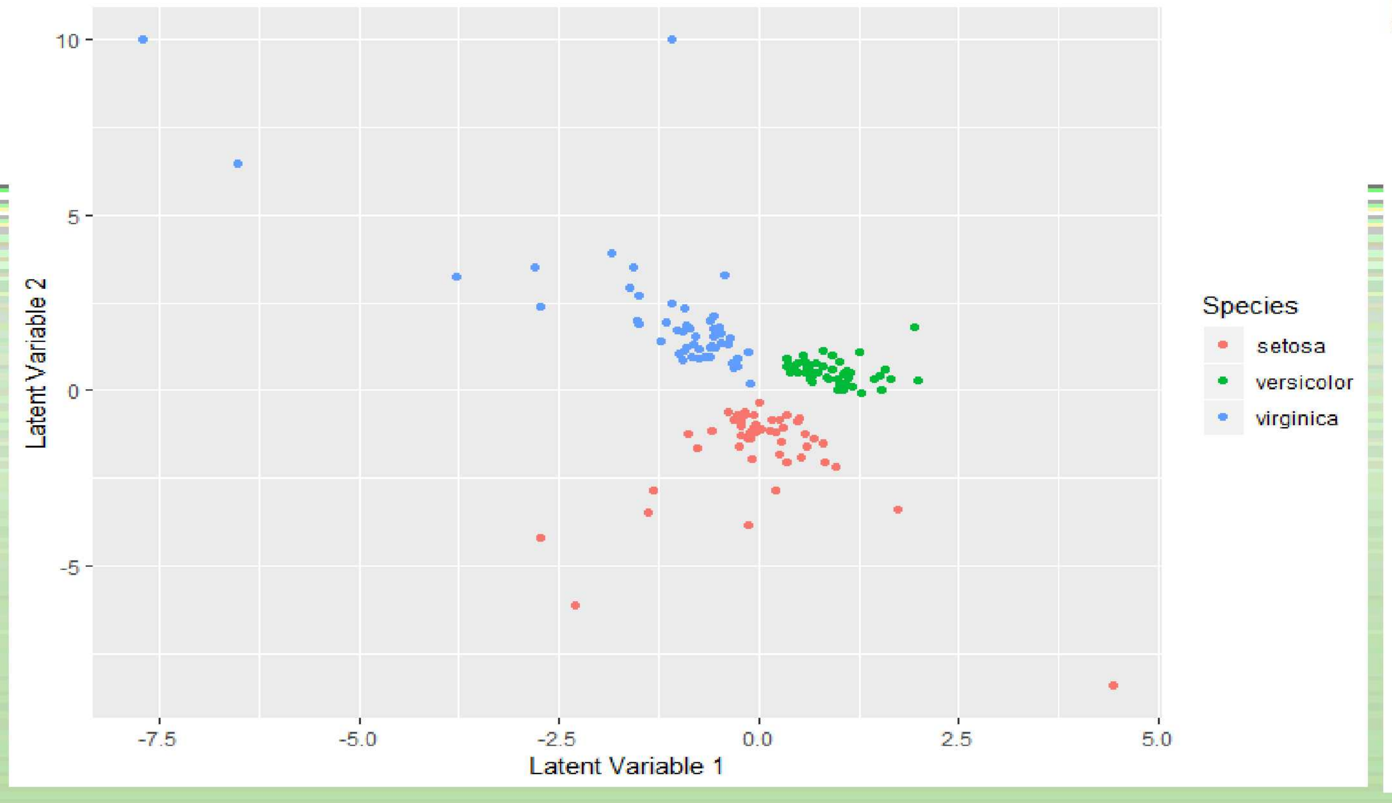
Summarize video output with aggregation method of choice:

- $P(at\ least\ 1\ object) = 1 - P(\text{no object})$
- $E[object]$
- Object Count: Hard-classifications

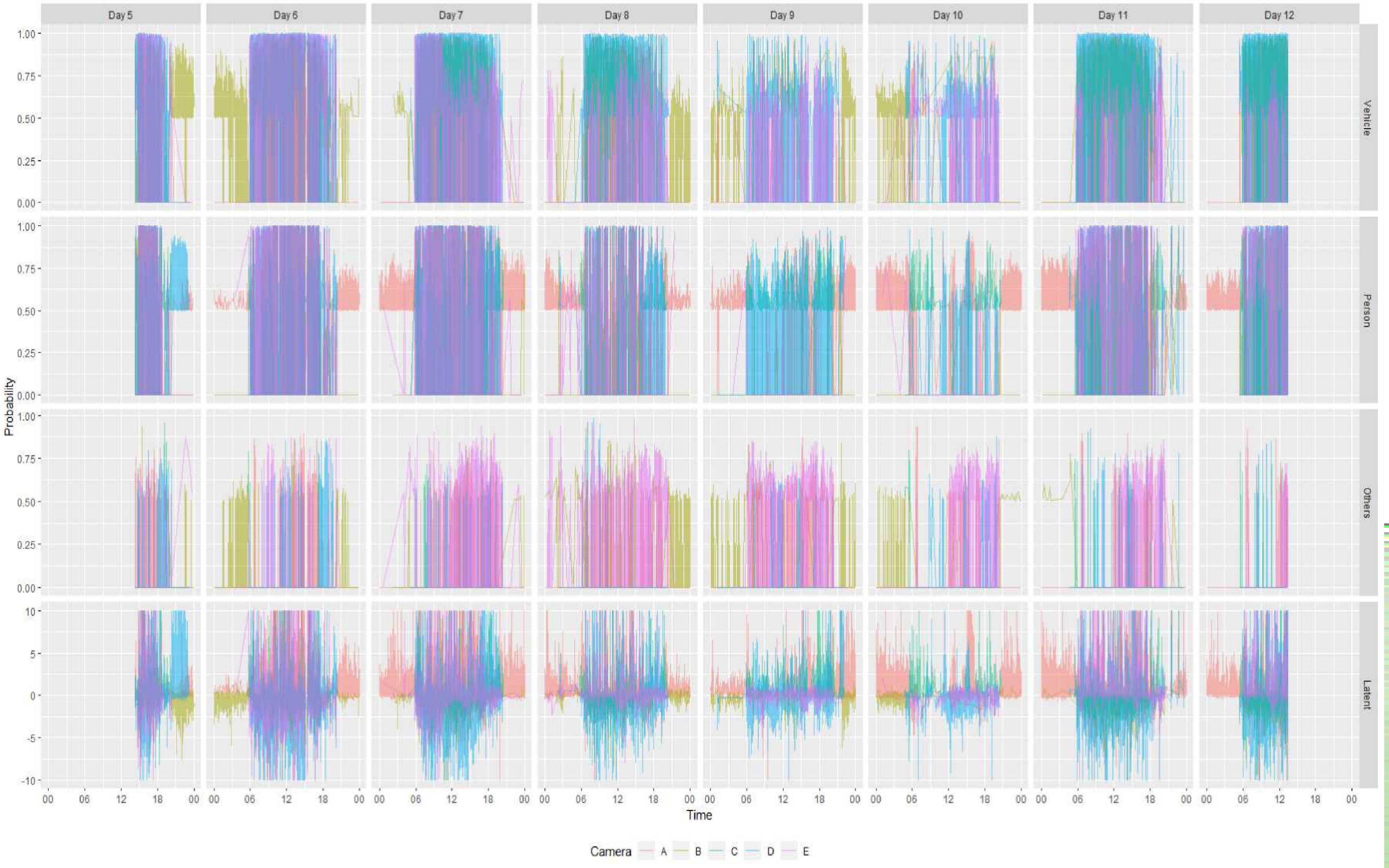
Combine with the weather data and fit using EM algorithm to get a latent activity variable

Challenges:

- Misclassification, coverage, and censoring of objects greatly affects downstream analysis
- Local optima: Multiple EM runs with random initializations used to explore non-convex likelihood function.
- Computational speed: C++ code for critical subfunctions, and parallel EM runs



(Above) Example of Tallis applied to the Iris data set to for dimension reduction.



(Above) using the output of a ML algorithm, probability of at least one of an object existing at a time point, plotted over time. Row one is vehicle, row 2 is people, row 3 is all other objects, row 4 is the result of our analysis. Colors denote different camera feeds.

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.
SAND No. ____________